FROM GEOMETRY TO GENERALIZATION: WHY ROW NORMALIZATION CAN BEAT ADAM AND MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Different optimizers can fit the same training data while selecting classifiers with substantially different geometries, but whether this difference provably affects population performance remains unclear. We show that row-wise normalization can achieve strictly higher population accuracy than full-batch Adam, a proxy for random-reshuffling Adam, and exact-SVD Muon in high-dimensional multi-class classification. Under an isotropic Gaussian-cloud data model, this advantage arises because row normalization's class-wise Euclidean geometry asymptotically preserves the population decision-boundary directions, whereas Adam's coordinate-wise geometry and Muon's spectral geometry introduce nonvanishing distortions. Beyond isotropy, the advantage persists for full-batch training on class means with independently oriented class-mean and test-noise covariances. It holds for power-law spectra with class-mean exponent below one, even under heavily anisotropic test noise. When both covariances are diagonal and sufficiently close, the advantage over Adam can reverse, while applying the same random rotation to both restores it by changing only their alignment with Adam's coordinate axes. Synthetic and last-layer language-model experiments support the predicted advantage.

1 INTRODUCTION

Row-wise normalization has recently emerged as a simple and effective ingredient in optimizers for large language model training.[1] SCALE (Glentis et al., 2025) uses row-normalized SGD as its base method and adds momentum only to the final layer, reporting performance comparable to or better than Adam (Kingma & Ba, 2015) and Muon (Jordan et al., 2024) while using substantially less optimizer-state memory.[2] Related row-level normalization operations also appear in recent methods such as SinkGD, NorMuon, RMNP, and MOGA (Scetbon et al., 2025; Li et al., 2025; Deng et al., 2026; Xu et al., 2026). Together, these results provide growing empirical evidence that normalizing updates separately across output units can be beneficial for large-scale optimization. Yet what drives the strong empirical performance of row-normalized methods remains poorly understood.

A natural hypothesis is that this behavior is tied to the geometry induced by the optimizer. Recent implicit-bias results make this connection concrete: normalized steepest-descent methods can converge toward max-margin classifiers associated with their underlying norm geometry (Soudry et al., 2018; Gunasekar et al., 2018a; Fan et al., 2025; Li et al., 2026; Baek et al., 2026). Under the implicit-bias characterizations used in this work, row-normalized methods are associated with the row norm[3], Muon with the spectral norm, and full-batch Adam with the entrywise max-norm. These distinct geometries can therefore lead the methods toward classifiers with substantially different decision boundaries, even when they all fit the training data well. What remains unclear is whether these geometric differences translate into systematic differences in population performance. This motivates our central question: *can row-wise normalization provably achieve higher population accuracy than competing optimizers, and if so, what geometric mechanism explains the advantage?*

Existing theory does not answer this question. The closest results analyze row-normalized methods through nonconvex convergence or width-scaling properties, but do not establish that they select statistically better solutions than competing optimizers (Deng et al., 2026; Xu et al., 2026).

[1] We use $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$; under the transposed convention, row-wise becomes column-wise normalization.
[2] SCALE uses the transposed weight convention and thus calls the same operation column-wise normalization.
[3] In our paper, row norm means $\|W\|_{\text{row}} := \max_i \|W_{i,:}\|_2$.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

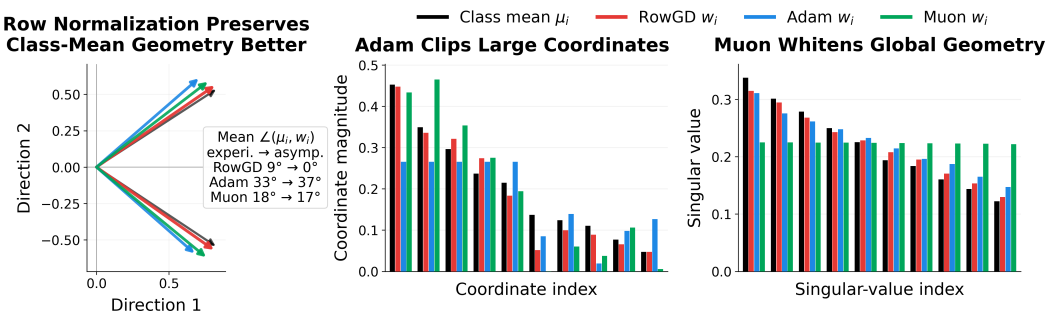

Figure 1: **Geometry of the learned classifiers.** Class means are compared with classifier rows learned by row-normalized GD, Adam, and Muon for 60 classes in 20 dimensions: a 2D projection of two classes, with mean angles over all classes and their proportional-limit predictions from Proposition 3.2 (left); coordinate magnitudes at ranks $1, 3, \ldots, 19$ after sorting by class-mean magnitude (middle); and singular values at indices $1, 3, \ldots, 19$ of the full matrices (right).

To the best of our knowledge, we provide the first such result. We analyze the statistical quality of the limiting classifiers selected by different optimizer geometries and establish a strict population-accuracy advantage for row-wise normalization in high-dimensional Gaussian data models. Our analysis also identifies the geometric mechanism behind this advantage: the class-wise Euclidean geometry of row-wise normalization better preserves the population decision-boundary directions, whereas coordinate-wise and spectral geometries introduce systematic directional distortions.

1.1 Our Contributions

We study how optimizer geometry affects population accuracy in high-dimensional multiclass classification. Our main contributions are the following.

- **Population-accuracy ordering under an isotropic Gaussian-cloud data model.** We first establish a strict population-accuracy advantage of row-normalized optimizers over representative spectral, coordinate-wise, and adaptive alternatives under our isotropic Gaussian-cloud data model. As an ingredient, we extend existing normalized steepest-descent implicit-bias analyses (Fan et al., 2025; Li et al., 2026) to the row norm. Specifically, we show the advantage in three settings as follows.
  - **Deterministic full-batch setting.** Row-normalized GD and row-normalized GD with momentum outperform Adam, SignGD, Signum, exact-SVD Muon, and Spectral-GD (Theorem 3.1).
  - **Stochastic random-reshuffling setting.** Under random reshuffling, row-normalized SGD with momentum outperforms stochastic Signum and stochastic exact-SVD Muon (Proposition 3.4).
  - **Random-reshuffling Adam via AdamProxy.** We extend the AdamProxy framework of Baek et al. (2026) from incremental Adam to random-reshuffling mini-batch Adam. We prove that the resulting proxy approximates the one-epoch displacement of random-reshuffling Adam in the $\beta_1, \beta_2 \to 1$ regime (Proposition 3.5) and show that row-normalized SGD with momentum achieves strictly higher population accuracy than AdamProxy (Proposition 3.6).
- **Extension beyond isotropy.** We show that the advantage of row-normalized methods is not limited to isotropic data distributions. For deterministic full-batch training on the class means, we consider class-mean and test-noise covariances with independent random eigenbases and power-law spectra, allowing any class-mean exponent below one and any test-noise exponent. Under these conditions, we prove that the row-normalized methods continue to outperform Adam and Muon (Theorem 4.2). We further exhibit a coordinate-aligned anisotropic regime in which row-normalized methods are worse than Adam, and show that jointly rotating the class means and test-noise covariance by an independently sampled Haar matrix restores the advantage with high probability (Proposition 4.5).

1.2 Related Work

Here we discuss only the works most directly related to our problem formulation and proof strategy; additional related work is deferred to Appendix A.

Row-normalized optimizers in practice. Recent LLM optimizers use output-wise row normalization as a principal matrix update in SCALE, RMNP, and MOGA; as one step of alternating row–column normalization in SinkGD; and as row-wise adaptive rescaling after Muon orthogonalization in NorMuon (Glentis et al., 2025; Scetbon et al., 2025; Li et al., 2025; Deng et al., 2026; Xu et al., 2026). These works establish row normalization as a practical optimizer component, whereas we theoretically compare the quality of the solutions it selects against competing optimizer geometries.

Theoretical comparisons involving row normalization. RMNP proves nonconvex convergence guarantees for row-normalized momentum methods that match existing guarantees for Muon (Deng et al., 2026). MOGA compares the width dependence of smoothness and learning-rate transfer across matrix geometries, showing favorable scaling for row-normalized methods (Xu et al., 2026).

Implicit bias and optimizer-selected solutions. Prior work characterizes the norm-induced max-margin directions of full-batch SignGD, Spectral-GD, Muon, and Adam in separable multiclass linear classification (Fan et al., 2025). This framework has been extended to random-reshuffling stochastic optimization, where batch size and momentum govern approximation to the full-batch margin (Li et al., 2026). Separately, Baek et al. (2026) show that incremental Adam can have a distinct stochastic implicit bias and introduce an AdamProxy characterized by a data-dependent Mahalanobis margin problem. We add the row norm and extend AdamProxy to multiclass random reshuffling.

2 PRELIMINARIES

Notation. For $k \geq 2$, let $[k] := \{1, \dots, k\}$ and $e_1, \dots, e_k$ be the standard basis of $\mathbb{R}^k$. For positive integers b, N, we write $b \mid N$ if N is divisible by b. For matrices, $\langle A, B\rangle := \mathrm{tr}(A^\top B)$, $\|A\|_{\max} := \max_{r,j} |A_{r,j}|$, and $\|A\|_{\mathrm{sum}} := \sum_{r,j} |A_{r,j}|$; we use $\|A\|_{\mathrm{op}}$ and $\|A\|_{\mathrm{F}}$ for the spectral and Frobenius norms. For $W \in \mathbb{R}^{k\times d}$ with rows $w_r^\top$, define $\|W\|_{\mathrm{row}} := \|W\|_{2,\infty} := \max_{r\in[k]} \|w_r\|_2$ as a row norm. For $a \in \mathbb{R}^k$ and $c \in [k]$, define $\mathbb{S}_c(a) := e^{a_c} / \sum_{\ell=1}^k e^{a_\ell}$, the c-th component of the softmax map. We write $A^{\odot p}$ for entrywise powers and $\oslash$ for entrywise division, with $0/0 := 0$. For nonzero vectors u, v, let $\angle(u, v) \in [0, \pi]$ denote the angle between u and v. Write $\widehat{v} := v/\|v\|_2$ for $v \neq \mathbf{0}$. For random quantities indexed by d, $o_{\mathbb{P}}(1)$ denotes a term that converges to zero in probability as $d \to \infty$.

Data model. We collect the structural assumptions on the data distribution in the following Gaussian-cloud data model; asymptotic and optimization conditions appear separately below. Let d be the input dimension, k the number of classes, and n_i the class-i sample size, $N := \sum_{i=1}^k n_i$ and $n_{\min} := \min_i n_i$.

Assumption 2.1 (Isotropic Gaussian-cloud data model)**.** *The training dataset consists of $(x_{i,s}, i)$ for $i \in [k]$ and $s \in [n_i]$ such that*

$$x_{i,s} := \mu_i + r z_{i,s}, \qquad \text{where} \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d), \qquad z_{i,s} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d). \tag{1}$$

A test example X^{te} with label Y is

$$X^{\mathrm{te}} := \mu_Y + r_{\mathrm{te}} z^{\mathrm{te}}, \qquad \mathbb{P}(Y = i) = \pi_{i,d}, \qquad z^{\mathrm{te}} \sim \mathcal{N}(\mathbf{0}, I_d),$$

where $(\mu_i)_{i=1}^k$, $(z_{i,s})_{i,s}$, Y, and z^{te} are mutually independent. Here, $r = r(d) \geq 0$, $r_{\mathrm{te}} = r_{\mathrm{te}}(d) > 0$, $\pi_{i,d} > 0$, and $\sum_{i=1}^k \pi_{i,d} = 1$.

We also use the flattened notation $\{(x_n, y_n)\}_{n=1}^N$ for the same training set, where $x_n = x_{i,s}$ and $y_n = i$ under some fixed bijection between $\{(i, s) : i \in [k],\, s \in [n_i]\}$ and $[N]$.

Assumption 2.2 (High-dimensional scaling)**.** *Along a sequence of models indexed by d, with $k = k(d)$, $n_i = n_i(d)$, and $\pi_{\min,d} := \min_i \pi_{i,d}$,*

$$k/d \to \alpha \in (0, \infty), \qquad n_{\min} \to \infty, \qquad \log N = o(d), \qquad \log(1/\pi_{\min,d}) = o(d). \tag{2}$$

Furthermore, $r_0 := \sup_d r(d) < \infty$ and $r_{\mathrm{te},0} := \sup_d r_{\mathrm{te}}(d) < \infty$.

We focus on a proportional regime in which k and d remain on the same scale, excluding settings where one asymptotically dominates the other; this is qualitatively compatible with language-model last layers. The condition $\log N = o(d)$ rules out datasets exponential in d, so that the geometry is not dominated by rare deviations accumulated over exponentially many samples. Meanwhile, $\log(1/\pi_{\min,d}) = o(d)$ allows class frequencies following Zipf's law, as observed in language data.

Classifier and loss. A linear multiclass classifier is $W \in \mathbb{R}^{k\times d}$, with score vector Wx at $x \in \mathbb{R}^d$. We train with multiclass cross-entropy

$$L(W) := -\frac{1}{N}\sum_{i=1}^{k}\sum_{s=1}^{n_i}\log \mathbb{S}_i(Wx_{i,s}).$$

This standard setting (Fan et al., 2025; Li et al., 2026) abstracts a language model's final linear layer.

Population accuracy. We measure population performance by the probability that the true class attains the strictly largest linear score on a test example:

$$\mathrm{Acc}_{\pi,d}(W) := \sum_{i=1}^{k}\pi_{i,d}\,\mathbb{P}\big(w_i^\top X^{\mathrm{te}} > w_j^\top X^{\mathrm{te}} \text{ for all } j \in [k]\setminus\{i\} \mid Y = i\big)\,.$$

Norm-based update maps. We use three norm-based update maps. For the row norm $\psi_{\mathrm{row}} := \|\cdot\|_{\mathrm{row}}$, set $(\Delta_{\mathrm{row}}(G))_{r,:} := G_{r,:}/\|G_{r,:}\|_2$ when $G_{r,:} \neq \mathbf{0}$, and $(\Delta_{\mathrm{row}}(G))_{r,:} := \mathbf{0}$ otherwise. For the entrywise max-norm $\psi_{\max} := \|\cdot\|_{\max}$, let $\Delta_{\max}(G) := \mathrm{sign}(G)$ entrywise, with $\mathrm{sign}(0) := 0$. For the spectral norm $\psi_{\mathrm{sp}} := \|\cdot\|_{\mathrm{op}}$, if $G \neq \mathbf{0}$, let $G = U\Sigma V^\top$ be a compact SVD and define $\Delta_{\mathrm{sp}}(G) := UV^\top$; set $\Delta_{\mathrm{sp}}(\mathbf{0}) := \mathbf{0}$. For each $g \in \{\mathrm{row}, \mathrm{sp}, \max\}$, these choices satisfy $\Delta_g(G) \in \arg\max_{\psi_g(\Delta)\le 1}\langle G, \Delta\rangle$, although the maximizer need not be unique.

Full-batch and random-reshuffling optimizers. For random reshuffling (RR), assume $b \mid N$, set $m := N/b$, and independently permute and partition the data into m mini-batches of size b each epoch. Let B_t be the batch at step t, define $L_{B_t}(W) := |B_t|^{-1}\sum_{(i,s)\in B_t} - \log \mathbb{S}_i(Wx_{i,s})$, and set $G_t := \nabla L(W_t)$ under full-batch training and $G_t := \nabla L_{B_t}(W_t)$ under RR. For $g \in \{\mathrm{row}, \mathrm{sp}, \max\}$, $\beta \in [0, 1)$, and $M_0 = \mathbf{0}$, we write

$$M_{t+1} = \beta M_t + (1-\beta)G_t, \qquad W_{t+1} = W_t - \eta_t \Delta_g(M_{t+1}).$$

Under full-batch training, $\beta = 0$ gives row-normalized gradient descent (RowGD), spectral gradient descent (SpecGD), and sign gradient descent (SignGD) for $g = \mathrm{row}, \mathrm{sp}, \max$, respectively; with $\beta \in (0, 1)$, the corresponding methods are row-normalized gradient descent with momentum (RowGDM), exact-SVD Muon (Muon),[4] and Signum. Under RR, we denote the momentum counterparts by RR-RowSGDM, RR-Muon, and RR-Signum. Thus, throughout the theory, Muon and RR-Muon refer to the exact-SVD variants. Throughout, W_t^{opt} denotes the parameter iterate of optimizer opt at step t.

Assumption 2.3 (Bounded number of mini-batches)**.** *Under random-reshuffling training with batch size $b = b_d$, the number of mini-batches per epoch, $m_d := N(d)/b_d$, satisfies $m_0 := \sup_d m_d < \infty$.*

Adam optimizer. We consider Adam without bias correction or a numerical-stability constant, with $0 < \beta_1 \le \beta_2 < 1$, using G_t from the corresponding full-batch or random-reshuffling setting:

$$M_{t+1} = \beta_1 M_t + (1-\beta_1)G_t, \quad V_{t+1} = \beta_2 V_t + (1-\beta_2)G_t^{\odot 2}, \quad W_{t+1} = W_t - \eta_t M_{t+1} \oslash V_{t+1}^{\odot 1/2}. \tag{3}$$

We initialize $M_0 = V_0 = \mathbf{0}$, and refer to the full-batch and random-reshuffling versions as Adam and RR-Adam, respectively. The following condition holds almost surely at $W_0 = 0$.

Assumption 2.4 (Full-batch Adam initialization)**.** *For full-batch Adam, every entry of $\nabla L(W_0)$ is nonzero.*

Learning-rate schedule. We use the following learning-rate conditions.

Assumption 2.5 (Base learning-rate condition)**.** *The sequence $\{\eta_t\}_{t\ge 0}$ is positive and nonincreasing, with $\eta_t \to 0$ and $\sum_{t=0}^{\infty}\eta_t = \infty$.*

Assumption 2.6 (Momentum learning rate)**.** *For every $\beta \in (0, 1)$ and $c_1 > 0$, there exist $t_0 \in \mathbb{N}$ and $c_2 > 0$ such that $\sum_{s=0}^{t}\beta^s\big(\exp(c_1\sum_{\tau=1}^{s}\eta_{t-\tau}) - 1\big) \le c_2\eta_t$ for all $t \ge t_0$.*

Non-momentum methods require only Assumption 2.5; momentum methods additionally require Assumption 2.6. Both hold for $\eta_t = c(t+1)^{-a}$, $c > 0$, $a \in (0, 1]$.

[4] Practical Muon (Jordan et al., 2024) approximates the orthogonalization step using a finite number of Newton–Schulz matrix-polynomial iterations. We instead use the exact SVD polar factor $UV^\top$; this is equivalent to adding momentum to SpecGD and gives the variant we refer to as exact-SVD Muon.

3 MAIN RESULTS

Section 3.1 establishes the deterministic optimizer population-accuracy ordering and identifies its limiting geometries. Section 3.2 studies robustness to random-reshuffling stochastic training, while Section 4 extends the analysis beyond the isotropic Gaussian-cloud data model.

3.1 COMPARISON OF OPTIMIZERS IN THE DETERMINISTIC SETTING

We first compare the full-batch optimizers defined in Section 2. Our first main result shows that RowGD and RowGDM achieve strictly higher population accuracy than SpecGD, Muon, SignGD, Signum, and Adam.

Theorem 3.1 (Deterministic optimizer accuracy ordering)**.** *Define the full-batch optimizer groups*

$$\mathcal{O}_{\mathrm{row}} := \{\mathrm{RowGD}, \mathrm{RowGDM}\}, \quad \mathcal{O}_{\mathrm{comp}} := \{\mathrm{SpecGD}, \mathrm{Muon}, \mathrm{SignGD}, \mathrm{Signum}, \mathrm{Adam}\}. \tag{4}$$

Suppose Assumptions 2.1, 2.2, 2.4, 2.5, and 2.6 hold. Then, for every sufficiently large d, with probability at least $1 - e^{-\sqrt{d}}$ over the training data, the following holds for all momentum parameters $\beta \in (0,1)$ of RowGDM, Muon, and Signum:

$$\min_{\mathrm{opt} \in \mathcal{O}_{\mathrm{row}}} \liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big) > \max_{\mathrm{opt} \in \mathcal{O}_{\mathrm{comp}}} \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big). \tag{5}$$

Proof idea. The proof proceeds through the implicit biases of the optimizers. RowGD and RowGDM approach row-norm margin solutions, SpecGD and Muon approach spectral-norm margin solutions, and SignGD, Signum, and Adam approach max-norm margin solutions. Here, a ψ-norm margin solution minimizes $\psi(W)$ among classifiers with training margin at least one; the formal solution sets are defined in Appendix C. Under the Gaussian-cloud model, these three solution sets have the distinct asymptotic classifier structures summarized in Proposition 3.2, and Gaussian pairwise-error bounds convert this geometric separation into the population-accuracy ordering. Because Theorem 3.1 is proven through a uniform comparison over all norm-induced margin solutions approached by the optimizers, it does not require directional convergence of optimizers.

Proposition 3.2 (Limiting classifier geometry)**.** *Let W_t^{opt} have rows $(w_{i,t}^{\mathrm{opt}})^\top$, and let $\Theta_t^{\mathrm{opt}} := k^{-1}\sum_{i=1}^k \angle(w_{i,t}^{\mathrm{opt}}, \mu_i)$ denote the average angle between the class mean and classifier row. Let $M_\mu \in \mathbb{R}^{k\times d}$ have rows $\mu_i^\top$, let $P := I_k - k^{-1}\mathbf{1}\mathbf{1}^\top$ and $r_d := \min\{k-1, d\}$, and let $C_\mu := PM_\mu$ have compact SVD $U_\mu \Sigma_\mu V_\mu^\top$ and polar factor $Q_\mu := U_\mu V_\mu^\top$. Suppose Assumptions 2.1, 2.2, 2.4, 2.5, and 2.6 hold. Then*

$$\begin{aligned}
\text{RowGD/RowGDM:} \quad & \limsup_{t\to\infty} \max_i \big\| \widehat{w}_{i,t}^{\mathrm{opt}} - \widehat{\mu}_i \big\|_2 = o_{\mathbb{P}}(1), \\
\text{SignGD/Signum/Adam:} \quad & \limsup_{t\to\infty} \max_i \left\| \widehat{w}_{i,t}^{\mathrm{opt}} - \frac{\mathrm{sign}(\mu_i)}{\sqrt{d}} \right\|_2 = o_{\mathbb{P}}(1), \\
\text{SpecGD/Muon:} \quad & \limsup_{t\to\infty} \frac{1}{r_d} \left\| \frac{PW_t^{\mathrm{opt}}}{\|PW_t^{\mathrm{opt}}\|_{\mathrm{op}}} - Q_\mu \right\|_{\mathrm{F}}^2 = o_{\mathbb{P}}(1).
\end{aligned}$$

Moreover, $\limsup_{t\to\infty} |\Theta_t^{\mathrm{opt}} - \theta| = o_{\mathbb{P}}(1)$, where $\theta = 0$ for RowGD/RowGDM and $\theta = \arccos\sqrt{2/\pi}$ for SignGD/Signum/Adam. If the aspect ratio $\alpha > 1$, then $\theta = \arccos\kappa_{\mathrm{sp}}(\alpha)$ for SpecGD/Muon, where $\kappa_{\mathrm{sp}} : (0,\infty) \to (0,1)$ is a positive constant depending only on α.[5]

Underlying mechanism. Proposition 3.2 makes the source of the accuracy gap explicit. RowGD and RowGDM preserve each class-mean direction, driving w_i toward a positive multiple of μ_i. In contrast, SignGD, Signum, and Adam approach the coordinate-wise sign direction of each class mean, effectively equalizing coordinate magnitudes. SpecGD and Muon approach the polar factor of the centered class-mean matrix, effectively equalizing its singular values. Both transformations introduce nonvanishing directional distortion relative to the original class-mean geometry. Under isotropic Gaussian test noise, preserving the class-mean geometry leads to better separation of unseen examples, while the distortions induced by the other methods degrade this separation. This yields the population-accuracy advantage of the row-normalized methods in Theorem 3.1.

[5] Its explicit formula is given in (41). Numerically, $\kappa_{\mathrm{sp}}(1) \approx 0.84883$ and $\kappa_{\mathrm{sp}}(2) \approx 0.93280$.

Corollary 3.3 (Population-error exponents). *Let* $\mathcal{E}_{t,d}^{\mathrm{opt}} := 1 - \mathrm{Acc}_{\pi,d}(W_t^{\mathrm{opt}})$. *Suppose Assumptions 2.1, 2.2, 2.4, 2.5, and 2.6 hold. Then*

$$\limsup_{t\to\infty} \left| \frac{r_{\mathrm{te}}^2}{d} \log \frac{\mathcal{E}_{t,d}^{\mathrm{Adam}}}{\mathcal{E}_{t,d}^{\mathrm{RowGDM}}} - \frac{\pi-2}{4\pi} \right| = o_{\mathbb{P}}(1), \quad \liminf_{t\to\infty} \frac{r_{\mathrm{te}}^2}{d} \log \frac{\mathcal{E}_{t,d}^{\mathrm{Muon}}}{\mathcal{E}_{t,d}^{\mathrm{RowGDM}}} \geq \frac{1-\kappa_{\mathrm{sp}}(\alpha)^2}{4} - o_{\mathbb{P}}(1),$$

where κ_{sp} *is the constant as in Proposition 3.2. The same conclusion holds with RowGDM replaced by RowGD, Adam by SignGD or Signum, and Muon by SpecGD.*

Corollary 3.3 follows from Proposition C.27 in Appendix C. While Theorem 3.1 establishes a strict ordering of the limiting population accuracies, Corollary 3.3 quantifies the separation. Although all three population errors vanish as $d \to \infty$, RowGDM has exponentially smaller population error than Adam and Muon. This separation is captured by the logarithms of the error ratios on the scale d/r_{te}^2.

3.2 Robustness to Stochastic Training

Mini-batch updates can alter the implicit bias of an optimizer, so the full-batch ordering does not automatically extend to stochastic training. Under RR, however, sufficiently large momentum keeps RR-RowSGDM, RR-Muon, and RR-Signum close to their respective row-, spectral-, and max-norm margin solutions (Li et al., 2026). We show that this approximate preservation is enough for the strict population-accuracy ordering to persist.

Proposition 3.4 (Stochastic optimizer accuracy ordering). *Suppose Assumptions 2.1, 2.2, 2.3, 2.5, and 2.6 hold. Then there exists* $\beta_0 = \beta_0(\alpha, r_0, m_0) \in [0,1)$ *such that, for every sufficiently large* d*, with probability at least* $1 - e^{-\sqrt{d}}$ *over the training data, the following holds for all momentum parameters* $\beta \in [\beta_0, 1)$ *of RR-RowSGDM, RR-Muon, and RR-Signum, and for every realization of the reshuffling:*

$$\liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\text{RR-RowSGDM}}\big) > \max_{\mathrm{opt}\in\{\text{RR-Muon, RR-Signum}\}} \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big). \tag{6}$$

The condition $m_d \leq m_0$ in Assumption 2.3 means that each mini-batch contains at least a $1/m_0$ fraction of the training set. For $m_0 \geq 2$, the sufficient momentum threshold used in the proof has $1-\beta_0$ proportional to $1/[m_0(m_0^2-1)]$.

Approximating Adam via AdamProxy. Adam requires separate treatment under RR because averaging squared mini-batch gradients generally differs from squaring the averaged gradient, so its full-batch max-norm margin characterization need not persist (Baek et al., 2026). To address this, we develop a deterministic proxy for RR-Adam in the multiclass setting with general mini-batch size, extending the binary incremental AdamProxy construction of Baek et al. (2026). Using the flattened indexing $\{(x_n, y_n)\}_{n=1}^N$, let $g_n(W) := -(e_{y_n} - \mathbb{S}(Wx_n))x_n^\top$ denote the gradient of the cross-entropy loss for sample (x_n, y_n). For each batch $B \subset [N]$ of size b, set $G_B(W) := b^{-1}\sum_{n\in B} g_n(W) = \nabla L_B(W)$. Thus, under RR, $G_t = G_{B_t}(W_t)$ in (3).

Drawing B uniformly from the b-subsets of $[N]$, define the AdamProxy direction

$$\Phi_{N,b}(W) := \mathbb{E}_B[G_B(W)] \oslash \mathbb{E}_B\big[G_B(W)^{\odot 2}\big]^{\odot 1/2}.$$

Equivalently, letting $S(W) := \sum_{n=1}^N g_n(W)$, $Q(W) := \sum_{n=1}^N g_n(W)^{\odot 2}$,

$$\Phi_{N,b}(W) = \sqrt{\frac{b(N-1)}{N}}\, S(W) \oslash \big((b-1)S(W)^{\odot 2} + (N-b)Q(W)\big)^{\odot 1/2}. \tag{7}$$

Proposition 3.5 shows that, in the $\beta_1, \beta_2 \to 1$ regime, this deterministic direction approximates the normalized one-epoch displacement of RR-Adam.

Proposition 3.5 (One-epoch $\beta_1, \beta_2 \to 1$ RR-Adam approximation). *For each epoch* $r = 0, 1, 2, \ldots$*, let* $\tau_r := mr$ *denote the index of its first update and* $\gamma_r := \sum_{s=0}^{m-1} \eta_{\tau_r+s}$ *its total learning rate. Fix a training set with* $x_n[a] \neq 0$ *for all* $n \in [N]$ *and* $a \in [d]$*. Suppose that* $b < N$ *and that Assumptions 2.5 and 2.6 hold. Then, for every* $\varepsilon > 0$*, RR-Adam in* (3) *satisfies*

$$\lim_{\substack{(\beta_1,\beta_2)\to(1,1)\\ 0<\beta_1\leq\beta_2<1}} \limsup_{r\to\infty} \mathbb{P}\left(\left\| \frac{W_{\tau_{r+1}}^{\text{RR-Adam}} - W_{\tau_r}^{\text{RR-Adam}}}{\gamma_r} + \Phi_{N,b}\big(W_{\tau_r}^{\text{RR-Adam}}\big) \right\|_{\max} > \varepsilon \right) = 0, \tag{8}$$

where the probability is over the random reshuffling.

Proposition 3.5 is a one-epoch $\beta_1, \beta_2 \to 1$ approximation at fixed data and dimension. It does not establish long-horizon agreement between the RR-Adam and AdamProxy trajectories.

Motivated by Proposition 3.5, let $\gamma_t := \sum_{s=0}^{m-1} \eta_{mt+s}$ and consider the AdamProxy dynamics

$$W_{t+1}^{\mathrm{AP}} = W_t^{\mathrm{AP}} - \gamma_t \Phi_{N,b}\big(W_t^{\mathrm{AP}}\big)\,. \tag{9}$$

Comparison of AdamProxy and RR-RowSGDM. Although AdamProxy need not approach the exact sign geometry of full-batch Adam, it retains a weaker coordinate-wise saturation effect. Once the classifier is scaled to have unit minimum training margin, every coordinate eventually stays within a constant multiple of the typical coordinate size (Proposition E.8). This forces a nonvanishing distortion from the class-mean geometry, whereas high-momentum RR-RowSGDM remains close to the class-mean geometry, yielding the population-accuracy ordering below.

Proposition 3.6 (RR-RowSGDM versus AdamProxy accuracy ordering)**.** *Suppose Assumptions 2.1, 2.2, 2.3, 2.5, and 2.6 hold. Then there exists $\beta_1 = \beta_1(\alpha, r_0, m_0) \in [0,1)$ such that, for every sufficiently large d, with probability at least $1 - e^{-\sqrt{d}}$ over the training data, the following holds for all momentum parameters $\beta \in [\beta_1, 1)$ of RR-RowSGDM and for every realization of the reshuffling:*

$$\liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\text{RR-RowSGDM}}\big) > \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{AP}}\big)\,.$$

4 Extensions to Anisotropic Gaussian Models

In Section 3, we studied an isotropic Gaussian-cloud data model in which both the class means and the within-class noise are isotropic. In this section, we examine when the advantage of row-normalized methods persists or reverses beyond isotropy. Motivated by the nontrivial covariance structure that can arise in language-model representations, we extend the analysis to anisotropic Gaussian distributions.

To simplify the analysis and cover a broader range of anisotropy, we consider clean training, with no within-class noise in the training data, under the following common data model.

Assumption 4.1 (Anisotropic clean Gaussian model)**.** *Let $\Gamma, \Sigma_1, \ldots, \Sigma_k$ be positive-definite covariance matrices with $\operatorname{tr}\Gamma = \operatorname{tr}\Sigma_i = d$ for every $i \in [k]$.*[6] *Training uses the clean class signals $\{(\mu_i, i)\}_{i=1}^k$. A test example with label Y is $X^{\mathrm{te}} := \mu_Y + r_{\mathrm{te}} z^{\mathrm{te}}$, where, conditionally on the covariance matrices, (Y, z^{te}) is independent of $(\mu_i)_{i=1}^k$ and*

$$\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Gamma), \qquad \mathbb{P}(Y = i) = \pi_{i,d} > 0, \qquad z^{\mathrm{te}} \mid Y \sim \mathcal{N}(\mathbf{0}, \Sigma_Y).$$

Along the sequence indexed by d, $k/d \to \alpha \in (0,\infty)$ and $r_{\mathrm{te}} > 0$ satisfies $r_{\mathrm{te},0} := \sup_d r_{\mathrm{te}}(d) < \infty$.

4.1 Persistence of Row-Norm Advantage under Independent Orientations

We first specialize Assumption 4.1 to trace-normalized power-law covariances with independent Haar eigenbases. Suppressing dimension subscripts, define $T_\vartheta := \sum_{a=1}^d a^{-\vartheta}$ and $D_\vartheta := \operatorname{diag}(1, 2^{-\vartheta}, \ldots, d^{-\vartheta})$. For $s_\mu, s_\Sigma \geq 0$, set

$$\Gamma := \frac{d}{T_{s_\mu}} U_\mu D_{s_\mu} U_\mu^\top, \qquad \Sigma_i := \frac{d}{T_{s_\Sigma}} O_i D_{s_\Sigma} O_i^\top. \tag{10}$$

Here $U_\mu, O_1, \ldots, O_k$ are mutually independent Haar matrices on $\mathsf{O}(d)$; the class-mean draws are independent of $(O_i)_{i=1}^k$, and the optimizer initializations are fixed and do not depend on $(O_i)_{i=1}^k$.

Theorem 4.2 (Accuracy ordering under anisotropy with independent orientations)**.** *Let $\mathcal{O}_{\mathrm{row}}$ and $\mathcal{O}_{\mathrm{comp}}$ be as in* (4)*, with fixed $\beta \in (0,1)$ for RowGDM, Muon, and Signum. Suppose Assumptions 4.1, 2.4, 2.5, and 2.6 hold, with the covariances in* (10) *for $0 \leq s_\mu < 1$ and $s_\Sigma \geq 0$. Assume that $\log(1/\pi_{\min,d}) = o(d/\log d)$ and that $W_t^{\mathrm{opt}}/\|W_t^{\mathrm{opt}}\|_{\mathrm{F}}$ converges almost surely as $t \to \infty$ for every $\mathrm{opt} \in \mathcal{O}_{\mathrm{row}} \cup \mathcal{O}_{\mathrm{comp}}$. Then there exist constants $c, C > 0$, independent of d, such that, for every sufficiently large d, with probability at least $1 - Cd^{-c}$ over U_μ, $O_1, \ldots, O_k$, and the class means,*

$$\min_{\mathrm{opt}\in\mathcal{O}_{\mathrm{row}}} \liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big) > \max_{\mathrm{opt}\in\mathcal{O}_{\mathrm{comp}}} \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big)\,. \tag{11}$$

[6] A fixed trace preserves total variance while allowing covariance spectra and orientations to vary.

Here c, C can be chosen independently of s_μ and s_Σ, although the sufficiently-large-d threshold may depend on them. In particular, under independent test-noise orientations, the advantage of row-normalized methods persists for arbitrary test-noise anisotropy $s_\Sigma \geq 0$.

Remark 4.3 (Role of directional convergence)**.** *Unlike the accuracy comparisons in Section 3, which hold uniformly over all classifier families that the normalized iterates approach and therefore do not require directional convergence, Theorem 4.2 assumes directional convergence for each optimizer. Related directional-convergence assumptions are common in the implicit-bias literature (Gunasekar et al., 2018b; Nacson et al., 2019a; Chizat & Bach, 2020; Baek et al., 2026). Under clean training, conditional on the training data, the limit of $W_t^{\mathrm{opt}}/\|W_t^{\mathrm{opt}}\|_{\mathrm{F}}$ is fixed and independent of the test-noise eigenbases, which enter only at test time. Hence, the concentration bounds over the Haar-random test-noise eigenbases apply to the fixed limiting direction.*

Structural intuition: effective signal dimension. The condition $s_\mu < 1$ keeps the class-distinguishing signal sufficiently spread across dimensions, so different class means have weak pairwise overlap. This allows the row-norm margin solutions to remain close to the corresponding class-mean directions, while satisfying the multiclass margin constraints. If the signal is instead concentrated in only a few directions, some class means can have small pairwise angles, and margin constraints can push the row-norm margin solutions away from the class-signal-preserving geometry.

Remark 4.4 (Connection to language-model representations)**.** *We find that the fitted class-mean spectral exponents can be smaller than one ($s_\mu < 1$) in language-model representations; see Appendix H.3. This provides a possible connection between the anisotropic regime covered by Theorem 4.2 and representation geometry in language models.*

4.2 Coordinate-aligned Reversal and Recovery under Haar Rotation

We next show that the advantage of row-normalized methods can reverse when the class-mean and test-noise covariances are diagonal and sufficiently close, but is restored after applying the same independently sampled Haar rotation to the class means and test-noise covariance.

The underlying issue is the interaction between signal and test-noise geometry. With a common test-noise covariance Σ, the normal of the Bayes boundary between classes i and j is proportional to $\Sigma^{-1}(\mu_i - \mu_j)$. Thus, if directions that strongly separate the class means also have large test variance, the Bayes boundary can differ substantially from the raw class-mean difference that the row-norm margin solution tends to preserve. The construction below places this signal–noise alignment along the coordinate axes used by Adam's entrywise normalization, where Adam's tendency to equalize coordinate magnitudes can reduce the influence of such high-variance coordinates.

Let the class-mean covariance be diagonal and let the common test-noise covariance interpolate between isotropic noise and the same covariance profile:

$$\Gamma_d := \mathrm{diag}(\gamma_{1,d}, \ldots, \gamma_{d,d}), \qquad \Sigma_{\lambda,d} := (1-\lambda) I_d + \lambda \Gamma_d, \qquad \lambda \in [0,1]. \tag{12}$$

Thus, $\Sigma_{\lambda,d}$ interpolates between isotropic test noise at $\lambda = 0$ and the class-mean covariance at $\lambda = 1$.

Proposition 4.5 (Coordinate dependence of the RowGDM–Adam ordering)**.** *Let $\lambda \in [0,1]$. Suppose that Assumptions 4.1, 2.4, 2.5, and 2.6 hold with $\Gamma = \Gamma_d$ and $\Sigma_1 = \cdots = \Sigma_k = \Sigma_{\lambda,d}$ as in* (12)*. Suppose also that $\log(1/\pi_{\min,d}) = o(d)$, $0 < \inf_{a,d} \gamma_{a,d} \leq \sup_{a,d} \gamma_{a,d} < \infty$, and the limits $\bar\gamma_q := \lim_{d\to\infty} d^{-1} \sum_{a=1}^d \gamma_{a,d}^q$ exist $\forall q \in \{1/2, 2\}$ and satisfy $2\bar\gamma_{1/2}^2 \bar\gamma_2 > \pi$.[7] Then $\lambda_c := (\pi/(2\bar\gamma_{1/2}^2) - 1)/(\bar\gamma_2 - 1)$ lies in $(0,1)$, and the following hold.*[8]

(i) Diagonal covariance. *For every $\lambda \in [0,1] \setminus \{\lambda_c\}$, there exist $c_\lambda, C_\lambda > 0$ such that, for every sufficiently large d, with probability at least $1 - C_\lambda e^{-c_\lambda d}$ over the class means, for all $\beta \in (0,1)$,*

$$\liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{RowGDM}}\big) > \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{Adam}}\big) \qquad \text{if } \lambda < \lambda_c,$$

$$\liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{Adam}}\big) > \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{RowGDM}}\big) \quad \text{if } \lambda > \lambda_c.$$

[7] This requires a sufficiently dispersed spectrum. Corollary G.8 gives a sufficient condition.

[8] The same conclusions hold with RowGDM replaced by RowGD.

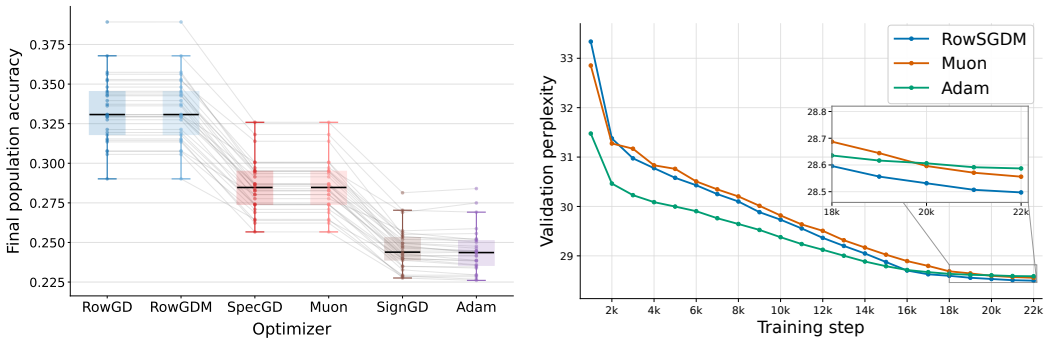

Figure 2: **Synthetic and language-model experiments. Left:** Final evaluation accuracy of six deterministic optimizers on 30 draws of Gaussian-cloud data. **Right:** Validation perplexity during last-layer-only LLaMA training with RowSGDM, Muon, and Adam. Lower perplexity is better.

(ii) Recovery under Haar rotation. *Let Q be Haar-uniformly random on $\mathsf{SO}(d)$ and independent of the class means, and replace every training and test input x by Qx. There exist $c_{\mathrm{rot}}, C_{\mathrm{rot}} > 0$ such that, for every fixed $\lambda \in [0, 1]$ and every sufficiently large d, with probability at least $1 - C_{\mathrm{rot}} e^{-c_{\mathrm{rot}} d}$ over the class means and Q, the RowGDM-over-Adam ordering holds.*

Implication. Theorem 4.2 shows that anisotropy alone does not eliminate the advantage of row-normalized methods under independent orientations. Proposition 4.5 identifies a structured failure mode. The RowGDM-over-Adam ordering can reverse when the high-variance class-mean and test-noise directions align with each other and with Adam's normalization coordinates. However, this reversal is coordinate-specific. Applying the same Haar rotation to the class means and test-noise covariance preserves the covariance spectra and their relative alignment while randomizing their orientation relative to the coordinate axes. This restores the RowGDM-over-Adam ordering. Hence, class-mean–test-noise alignment alone is not sufficient; the reversal additionally relies on its alignment with Adam's coordinate axes.

5 Experiments

Synthetic experiment. On isotropic Gaussian-cloud data with $k = 60$, $d = 20$, RowGD and RowGDM outperform SpecGD, Muon, SignGD, and Adam on all 30 paired seeds (Figure 2, left), consistent with Theorem 3.1. Additional comparisons and details are given in Appendix H.1.

Language model experiment. We freeze a trained LLaMA-60M model, zero-initialize its output head, and train only this layer with practical implementations of RowSGDM, Muon, or Adam. Figure 2 (right) shows that RowSGDM attains the lowest final validation perplexity, consistent with the qualitative ordering our theory predicts. Multi-seed results and details are given in Appendix H.2.

6 Conclusion

Implications. We connect optimizer-induced implicit-bias geometry to population accuracy in high-dimensional multiclass classification, beyond prior work on optimizer-selected implicit biases (Fan et al., 2025; Li et al., 2026; Baek et al., 2026). For linear classification abstracting the language model's last layer, row-normalized methods achieve strictly higher population accuracy than spectral and coordinate-wise alternatives in both full-batch and random-reshuffling settings by better preserving class-mean geometry. These results complement empirical evidence of competitive performance by row-normalized methods in language model training (Glentis et al., 2025; Deng et al., 2026).

Limitations and outlook. Our analysis is limited to a linear classification model that can be viewed as a last-layer readout, together with a stylized data model with Gaussian class means and within-class noise. Although the data model shares some high-dimensional geometric properties with linguistic collapse (Wu & Papyan, 2024), modern language-model representations are substantially more complex. This motivates extensions to richer representation models and multilayer training. Our anisotropic independent-orientation analysis assumes directional convergence, while our Adam results assume a zero numerical-stability constant and treat RR-Adam only via AdamProxy.

AI USE STATEMENT

In this work, we used generative AI tools to assist with improving English writing and readability, identifying potentially relevant related work that we might have overlooked, and drafting proof sections of the appendix by converting lengthy mathematical proofs and derivations into clean and readable LaTeX. We additionally used generative AI tools to assist with debugging experimental code and improving the visual presentation of figures and tables. All AI-assisted outputs were reviewed by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We state the assumptions and settings for our theoretical results in Sections 2 and 4, and provide complete proofs and supporting technical results in Appendices B–G. For the empirical results, Appendix H provides the experimental setups and information needed for reproduction, with synthetic experiments in Appendix H.1, the language-model experiment in Appendix H.2, and the representation analysis in Appendix H.3.

REFERENCES

Beomhan Baek, Minhak Song, and Chulhee Yun. Implicit bias of per-sample adam on separable data: Departure from the full-batch regime. In *International Conference on Learning Representations*, volume 2026, pp. 120180–120229, 2026.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34: 8407–8418, 2021.

Da Chang, Qiankun Shi, Lvgang Zhang, Yu Li, Ruijie Zhang, Yao Lu, Yongxiang Liu, and Ganzhao Yuan. Muoneq: Balancing before orthogonalization with lightweight equilibration. *arXiv preprint arXiv:2603.28254*, 2026.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pp. 1305–1338. PMLR, 2020.

Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

Alec Dewulf, Dhruv Pai, Li Yang, Ashley Zhang, and Ben Keigwin. Aurora: A leverage-aware spectral optimizer. *arXiv preprint arXiv:2606.27715*, 2026.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *Advances in Neural Information Processing Systems*, 38:39622–39669, 2025.

Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. Memory-efficient llm pretraining via minimalist optimizer design. *arXiv preprint arXiv:2506.16659*, 2025.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. *arXiv preprint arXiv:2602.16340*, 2026.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018b.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Jihwan Kim, Dogyoon Song, and Chulhee Yun. Scaling laws of signsgd in linear regression: When does it outperform sgd? In *International Conference on Learning Representations*, volume 2026, pp. 67140–67228, 2026a.

Jueun Kim, Baekrok Shin, Jihun Yun, Beomhan Baek, Minhak Song, and Chulhee Yun. Amuse: Anytime muon with stable gradient evaluation. *arXiv preprint arXiv:2605.22432*, 2026b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, volume 2015, 2015.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023.

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. *arXiv preprint arXiv:2602.11557*, 2026.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: Sgd with normalization and whitening enables stateless llm training. *arXiv preprint arXiv:2412.13148*, 2024.

Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pp. 6874–6883. PMLR, 2020.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pp. 62061–62121, 2025.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019b.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Qian Qian and Xiaoyuan Qian. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *Advances in neural information processing systems*, 37:81324–81359, 2024.

Meyer Scetbon, Chao Ma, Wenbo Gong, and Ted Meeds. Gradient multi-normalization for efficient llm training. *Advances in Neural Information Processing Systems*, 38:43001–43030, 2025.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Vignesh Subramanian, Rahul Arya, and Anant Sahai. Generalization for multiclass classification with overparameterized linear models. *Advances in Neural Information Processing Systems*, 35: 23479–23494, 2022.

Nikolaos Tsilivis, Eitan Gronich, Julia Kempe, and Gal Vardi. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. *Journal of Machine Learning Research*, 27 (104):1–37, 2026.

Sharan Vaswani, Reza Babanezhad, Jose Gallego-Posada, Aaron Mishkin, Simon Lacoste-Julien, and Nicolas Le Roux. To each optimizer a norm, to each norm its generalization. *arXiv preprint arXiv:2006.06821*, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices., 2012.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pp. 10849–10858. PMLR, 2021a.

Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35:26764–26776, 2022.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34: 24164–24179, 2021b.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. *Advances in Neural Information Processing Systems*, 37:137432–137473, 2024.

Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞ norm constrained optimization. *arXiv preprint arXiv:2404.04454*, 2024.

Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms i: Row/column normalization and hyperparameter transfer. *arXiv preprint arXiv:2603.09952*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jinghui Yuan, Jiaxuan Zou, Shuo Wang, Yong Liu, and Feiping Nie. Nora: Normalized orthogonal row alignment for scalable matrix optimizer. *arXiv preprint arXiv:2605.03769*, 2026.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in neural information processing systems*, 37:23988–24021, 2024.

OVERVIEW OF THE APPENDIX

The appendix is organized around the proof architecture of the paper. Appendix A provides additional related work, and Appendix B develops the optimizer-side margin and implicit-bias machinery for full-batch and random-reshuffling training. Appendices C–E combine this machinery with high-dimensional geometry to prove the deterministic, stochastic, and AdamProxy population-accuracy results. Appendix F extends the analysis to anisotropic Gaussian models with independently oriented test-noise covariances, Appendix G identifies a coordinate-specific RowGDM–Adam reversal and its recovery under a common Haar rotation, and Appendix H provides the experimental details and additional empirical evidence.

(1) In Appendix A, we provide additional related work on implicit bias in separable classification, adaptive and sign-based optimization, optimizer geometry and population performance, matrix-aware optimization geometry, row normalization and row–spectral hybrids, high-dimensional classification under mixture models, and last-layer geometry and neural collapse.

(2) In Appendix B, we establish the optimizer-side margin-convergence results used throughout the paper. We extend full-batch normalized steepest-descent and momentum arguments to max–sum dominated matrix norms and establish the max-norm maximum-margin characterization of full-batch Adam under our learning-rate assumptions. For random-reshuffling momentum methods, we develop a geometry-adapted analysis through the primal–dual atom radius $R_\psi(X)$, which applies directly to the Row, Spectral, and Max geometries. Specializing these results to row normalization yields the exact full-batch RowGD and RowGDM direction-set convergence in Proposition B.14 and the RR-RowSGDM near-margin guarantee in Proposition B.15.

(3) In Appendix C, we prove the deterministic population-accuracy ordering in Theorem 3.1 uniformly over the complete Row, Spectral, and Max maximum-margin direction sets. The proof establishes the all-aspect-ratio geometry of these three families: Row representatives align uniformly with the population mean differences, Max representatives exhibit a nonvanishing clipped-coordinate angular gap, and Spectral representatives are controlled through the reduced polar geometry of the centered empirical mean matrix, including the hard-edge regime at $\alpha = 1$. We then develop a geometry-to-population-error transfer and derive quantitative population-error exponents for the three geometries. The same exponent analysis also yields the full-batch Muon–Adam comparison in Corollary C.28.

(4) In Appendix D, we prove the random-reshuffling stochastic ordering in Proposition 3.4. The geometry-adapted momentum result from Appendix B places RR-RowSGDM, RR-Muon, and RR-Signum in near-maximum-margin direction sets for the Row, Spectral, and Max geometries. We reduce these near-margin directions to near-minimum-norm representatives, prove quantitative stability of the three near-minimizer families, and obtain a dimension-independent momentum threshold under which the deterministic geometric separation persists. The population-error transfer from Appendix C then gives the claimed stochastic accuracy ordering.

(5) In Appendix E, we analyze RR-Adam and AdamProxy. We first prove the one-epoch $\beta_1, \beta_2 \to 1$ RR-Adam approximation in Proposition 3.5; this result connects RR-Adam to the deterministic proxy over one epoch and does not assert long-horizon agreement of their trajectories. For AdamProxy itself, we establish positive margin growth and a trajectory-level coordinate cap, use the resulting coordinate saturation to obtain a nonvanishing boundary-direction gap, and combine this gap with the high-momentum RR-RowSGDM geometry and the population-error transfer from Appendix C to prove Proposition 3.6 without directional convergence. Separately, Proposition E.10 characterizes the limiting AdamProxy classifier through a self-consistent semidefinite quadratic margin problem when the normalized proxy direction converges.

(6) In Appendix F, we prove the anisotropic accuracy ordering in Theorem 4.2 under its directional-convergence assumption. Because training is clean, each optimizer-selected limiting direction

depends only on the class means and is independent of the Haar eigenbases of the test-noise covariances. We characterize the selected Row, Spectral, and Max directions, expose the independent test-noise orientations, and compare their standardized Gaussian test signals. This establishes the Row advantage throughout $0 \leq s_\mu < 1$ and $s_\Sigma \geq 0$, including arbitrarily strong test-noise anisotropy.

(7) In Appendix G, we prove the coordinate-specific RowGDM–Adam reversal in Proposition 4.5. Working uniformly over the complete Row and Max maximum-margin solution sets, we derive an explicit transition point λ_c for diagonal class-mean covariance with aligned anisotropic test noise, below which Row dominates and above which the ordering reverses. We then show that applying a common independent Haar rotation to the same class-mean and test-noise geometry restores the Row-over-Max, and hence RowGDM-over-Adam, ordering while preserving the rotation-invariant geometry. We also give a sufficient spectral-dispersion condition under which the transition regime is nonvacuous.

(8) In Appendix H, we provide the full experimental setup and additional empirical evidence. This includes expanded deterministic and random-reshuffling synthetic optimizer comparisons, auxiliary optimizer geometries, robustness studies across data and noise settings, synthetic experiments for the anisotropic independent-orientation and coordinate-aligned predictions, controlled last-layer language-model experiments with RowSGDM, Muon, and Adam, and measurements of class-mean and within-class covariance spectral exponents in language-model representations.

A Additional Related Works

Implicit bias in separable classification. The modern implicit-bias literature begins with the observation that, on linearly separable data, gradient descent on unregularized logistic or cross-entropy loss diverges in norm while converging in direction to a hard-margin classifier (Soudry et al., 2018). More generally, the geometry of the update rule can determine the norm with respect to which the limiting margin is maximized, as formalized for steepest-descent and mirror-descent methods by Gunasekar et al. (2018a). For Euclidean methods, this picture has been extended from gradient descent to fixed-step stochastic gradient descent and momentum methods (Nacson et al., 2019b; Wang et al., 2022), and recent work gives multiclass convergence results for broad families of exponentially tailed losses (Ravi et al., 2024). In homogeneous neural networks, norm-dependent steepest-descent dynamics generally converge to first-order stationary points of the corresponding margin problem rather than necessarily to its global optimum (Tsilivis et al., 2026), and a recent extension treats normalized and momentum variants of SignGD, Muon, and Adam (Gronich & Vardi, 2026). The results closest to ours characterize the max-margin directions selected by entrywise and Schatten geometries in multiclass linear models, their approximation under random reshuffling, and the data-dependent direction selected by incremental Adam (Fan et al., 2025; Li et al., 2026; Baek et al., 2026). Our analysis complements these directional results by adding the row norm and by comparing the population accuracy of entire optimizer-selected margin faces rather than only identifying a limiting direction.

Adaptive and sign-based optimization. Adam combines momentum with coordinate-wise second-moment normalization (Kingma & Ba, 2015), whereas signSGD and Signum retain only coordinate-wise gradient signs, with or without momentum (Bernstein et al., 2018). This relationship is not merely cosmetic: Adam can be decomposed into a sign direction and a variance-dependent magnitude, which helps explain both its robustness and its sensitivity to the coordinate system (Balles & Hennig, 2018); signSGD can also exhibit distinct compute-optimal scaling behavior from SGD (Kim et al., 2026a). The asymptotic classifier also depends on algorithmic details that are often suppressed in informal comparisons. For example, AdaGrad can converge to a margin solution determined by its path-dependent diagonal conditioner (Qian & Qian, 2019), adaptive methods with a positive stability constant can become asymptotically Euclidean in homogeneous models (Wang et al., 2021a), and Adam without a stability constant selects an ℓ_∞-max-margin direction in binary linear classification (Zhang et al., 2024). Decoupled weight decay changes the problem again: the implicit regularization of AdamW is related to an ℓ_∞-constrained objective rather than to the unregularized margin limit (Loshchilov & Hutter, 2017; Xie & Li, 2024). Accordingly, our full-batch Adam statement concerns the zero-stability-constant, no-weight-decay dynamics used in our theoretical model, while our

AdamProxy analysis isolates the additional data dependence created by incremental or mini-batch updates under random reshuffling (Baek et al., 2026).

Optimizer geometry and population performance. Implicit-bias results do not by themselves imply that one optimizer generalizes better than another. Constructed separable problems show that Euclidean gradient methods and adaptive coordinate-wise methods can interpolate the same training data yet have sharply different test errors (Wilson et al., 2017), and norm-based analyses relate such differences to the interaction between an optimizer's geometry and the geometry of the data (Vaswani et al., 2020). Experiments on transformers further suggest that the sign-like component of Adam, rather than stochastic noise alone, can account for part of its difference from SGD (Kunstner et al., 2023). Our contribution is a distribution-specific, high-probability ordering of row, spectral, and entrywise geometries in a multiclass model, uniform over the corresponding complete margin faces.

Matrix-aware optimization geometry. Several optimizer families exploit matrix or tensor structure without implementing the normalized steepest maps studied here. Shampoo uses Kronecker-factored, mode-wise preconditioners for tensor parameters (Gupta et al., 2018), while SOAP applies Adam in an evolving Shampoo eigenbasis (Vyas et al., 2025). The modular-duality framework instead assigns layerwise operator norms from the semantics of a module and derives corresponding duality maps (Bernstein & Newhouse, 2024), and norm-constrained linear minimization oracles provide a related route to architecture-aware updates (Pethick et al., 2025). Muon can be viewed as the matrix duality map obtained from the polar factor, commonly approximated by Newton–Schulz iterations (Jordan et al., 2024; Bernstein & Newhouse, 2024). By contrast, our spectral comparator is the exact polar update $UV^\top$ for a gradient with compact singular value decomposition $U\Sigma V^\top$; our theorems do not automatically transfer to preconditioned methods such as Shampoo or SOAP, or to finite-iteration approximations of the polar factor.

Row normalization and row–spectral hybrids. The row-structured methods summarized in Section 1.2 are algorithmically distinct even when they contain a superficially similar normalization step. In particular, SinkGD alternates row and column normalization, NorMuon performs neuron-wise adaptation after Muon orthogonalization, RMNP analyzes row-normalized momentum through nonconvex convergence guarantees, and MOGA studies width scaling and hyperparameter transfer for the corresponding operator geometry (Scetbon et al., 2025; Li et al., 2025; Deng et al., 2026; Xu et al., 2026). Recent hybrid proposals further modify the polar update by projecting momentum before row alignment, equilibrating rows before orthogonalization, or imposing leverage-aware row uniformity while retaining spectral structure (Yuan et al., 2026; Chang et al., 2026; Dewulf et al., 2026). SWAN is also a relevant stateless optimizer that combines column normalization with whitening (Ma et al., 2024). These methods strengthen the practical motivation for studying row geometry, but the extra projection, whitening, equilibration, or orthogonalization steps make them different dynamical systems. Our results should therefore be read as an analysis of exact row-normalized steepest descent and momentum, and as a controlled comparison that isolates the row normalization component rather than as an implicit-bias theorem for every row–spectral hybrid.

High-dimensional classification under mixture models. Gaussian and sub-Gaussian mixture models have been used to obtain sharp asymptotic formulas or risk bounds for regularized convex classifiers and binary max-margin classifiers (Mignacco et al., 2020; Cao et al., 2021). In the multiclass setting, prior work relates the hard-margin multiclass SVM to minimum-norm interpolation and studies benign overfitting under Gaussian designs (Wang et al., 2021b), including regimes in which the number of classes grows with dimension (Subramanian et al., 2022). These works primarily fix an estimator geometry, usually Euclidean, and characterize when its interpolation or maximum-margin bias generalizes. We instead hold the data model fixed and compare the distinct margin geometries selected by several optimizers in the proportional regime $k/d \to \alpha \in (0, \infty)$. The uniform statements over complete margin faces are important in this regime because a norm-specific margin problem need not have a unique optimizer, so identifying one convenient solution would not establish an optimizer-level population comparison.

Last-layer geometry and neural collapse. The class-conditional mean geometry in our model is also related to neural collapse, where terminal-phase representations approach within-class concentration, centered class means form a simplex-like frame, and classifier weights align with those means

(Papyan et al., 2020). Generalized neural-collapse formulations extend this viewpoint to settings with many more classes than feature dimensions by replacing the simplex equiangular tight frame with a one-versus-rest margin criterion (Jiang et al., 2023). Related geometric regularities have been reported for the unembedding representations of language models under the name linguistic collapse (Wu & Papyan, 2024). Our model does not attempt to explain the feature-learning dynamics that create such configurations. Instead, it conditions on a fixed last-layer representation with random Gaussian class centers and isotropic or anisotropic residual variation, and asks which optimizer-induced norm produces the most accurate off-sample decision regions once that representation has been formed.

B Implicit Bias of Row-Normalized Methods: Full-Batch and Random Reshuffling

This appendix collects the optimizer-side implicit-bias results used throughout the paper. For full-batch normalized steepest descent, we extend the analysis of Fan et al. (2025) from entrywise and Schatten norms to every max–sum dominated matrix norm. For random-reshuffling momentum methods, we extend the analysis of Li et al. (2026) and use a sharper geometry-adapted analysis based on the primal–dual atom radius $R_\psi(X)$, which avoids the common entrywise ℓ_1 feature bound and applies directly to the Row, Spectral, and Max geometries. We then specialize these general results to row normalization, thereby proving Proposition B.14 and the RR-RowSGDM margin result used in the stochastic accuracy theorem.

B.1 Common setup and auxiliary lemmas

Let $X = \{(x_n, y_n)\}_{n=1}^N$ be a fixed linearly separable multiclass dataset with $x_n \in \mathbb{R}^d$ and $y_n \in [k]$. A linear classifier is $W \in \mathbb{R}^{k\times d}$, and the cross-entropy loss is

$$L(W) := \frac{1}{N}\sum_{n=1}^{N} \ell_n(W), \qquad \ell_n(W) := -\log \mathbb{S}(Wx_n)_{y_n}.$$

Define

$$\mathrm{mar}_X(W) := \min_{n\in[N]} \min_{c\neq y_n} (w_{y_n} - w_c)^\top x_n, \qquad \gamma_\psi(X) := \max_{\psi(W)\le 1} \mathrm{mar}_X(W)$$

for any matrix norm ψ, and define the standard cross-entropy proxy

$$\mathcal{G}(W) := \frac{1}{N}\sum_{n=1}^{N} \big(1 - \mathbb{S}(Wx_n)_{y_n}\big).$$

For $0 \le \varepsilon < \gamma_\psi(X)$, define the near-max-margin direction set

$$\mathcal{U}^X_{\psi,\varepsilon} := \left\{ W \neq \mathbf{0} : \frac{\mathrm{mar}_X(W)}{\psi(W)} \ge \gamma_\psi(X) - \varepsilon \right\}, \qquad \mathcal{U}^X_\psi := \mathcal{U}^X_{\psi,0}.$$

For a nonzero matrix A and a nonempty set $\mathcal{S} \subset \mathbb{R}^{k\times d} \setminus \{\mathbf{0}\}$, let

$$d_{\cos}(A, \mathcal{S}) := 1 - \sup_{V\in\mathcal{S}} \frac{\langle A, V\rangle}{\|A\|_{\mathrm{F}}\|V\|_{\mathrm{F}}}.$$

Lemma B.1 (Effective margin implies direction-set convergence)**.** *Suppose $W_t \neq \mathbf{0}$ eventually and, for some $0 \le \varepsilon < \gamma_\psi(X)$,*

$$\liminf_{t\to\infty} \frac{\mathrm{mar}_X(W_t)}{\psi(W_t)} \ge \gamma_\psi(X) - \varepsilon. \tag{13}$$

Then

$$d_{\cos}(W_t, \mathcal{U}^X_{\psi,\varepsilon}) \longrightarrow 0.$$

Proof. By scale invariance of $d_{\cos}$, consider $W_t/\psi(W_t)$, which lies on the compact ψ-unit sphere. If the conclusion failed, a subsequence whose cosine discrepancy from $\mathcal{U}^X_{\psi,\varepsilon}$ is bounded below by a positive constant would have a further subsequence converging to some U_∞. Continuity of the margin and (13) give $\mathrm{mar}_X(U_\infty) \ge \gamma_\psi(X) - \varepsilon$, so $U_\infty \in \mathcal{U}^X_{\psi,\varepsilon}$, contradicting continuity of the cosine discrepancy. □

Lemma B.2 (Exact margin implies direction-set convergence)**.** *If*

$$\frac{\mathrm{mar}_X(W_t)}{\psi(W_t)} \longrightarrow \gamma_\psi(X),$$

then

$$d_{\cos}(W_t, \mathcal{U}_\psi^X) \longrightarrow 0.$$

Proof. Apply Lemma B.1 with $\varepsilon = 0$. □

For a learning-rate sequence, write $S_t := \sum_{s=0}^{t-1} \eta_s$.

Lemma B.3 (Cumulative remainders for general learning rates)**.** *Suppose Assumption 2.5 holds. Then $S_t \to \infty$, and for every $q > 1$, $c > 0$, and fixed $t_0 \in \mathbb{N}$,*

$$\frac{\sum_{s=0}^{t-1} \eta_s^q}{S_t} \longrightarrow 0, \qquad \sum_{s=t_0}^{t-1} \eta_s \exp\left(-c \sum_{\tau=t_0}^{s-1} \eta_\tau\right) = O(1).$$

Proof. Fix $\varepsilon > 0$ and choose T so that $\eta_s^{q-1} \le \varepsilon$ for all $s \ge T$. Then

$$\frac{\sum_{s=0}^{t-1} \eta_s^q}{S_t} \le \frac{\sum_{s=0}^{T-1} \eta_s^q}{S_t} + \varepsilon,$$

which proves the first limit. For the second assertion, enlarge $T \ge t_0$ so that $c\eta_s \le 1$ for $s \ge T$, and set $A_s := \sum_{\tau=t_0}^{s-1} \eta_\tau$. Since $1 - e^{-x} \ge x/2$ for $x \in [0, 1]$,

$$e^{-cA_s} - e^{-cA_{s+1}} \ge \frac{c}{2} \eta_s e^{-cA_s}.$$

The tail therefore telescopes to a bounded quantity, while the finite prefix is harmless. □

Lemma B.4 (Normalization of a cumulative margin bound)**.** *Let $\mathfrak{m} : \mathbb{R}^{k\times d} \to \mathbb{R}$ be continuous and positively homogeneous, let $\Gamma_* := \max_{\psi(W)\le 1} \mathfrak{m}(W) > 0$, and suppose $0 < \Gamma \le \Gamma_*$. Assume $S_t \to \infty$. If*

$$\mathfrak{m}(W_t) \ge \Gamma S_t - r_t, \qquad \psi(W_t) \le C_0 + S_t, \qquad r_t = o(S_t),$$

then

$$\liminf_{t\to\infty} \frac{\mathfrak{m}(W_t)}{\psi(W_t)} \ge \Gamma.$$

If $\Gamma = \Gamma_*$, then $\mathfrak{m}(W_t)/\psi(W_t) \to \Gamma_*$, and every accumulation point of $W_t/\psi(W_t)$ maximizes $\mathfrak{m}$ over the ψ-unit ball.

Proof. Since $r_t = o(S_t)$ and $S_t \to \infty$, the lower bound $\Gamma S_t - r_t$ is positive for all sufficiently large t. In particular, $W_t \ne 0$ eventually. For such t,

$$\frac{\mathfrak{m}(W_t)}{\psi(W_t)} \ge \frac{\Gamma S_t - r_t}{C_0 + S_t} \longrightarrow \Gamma.$$

This proves the liminf statement without imposing a sign condition on r_t. When $\Gamma = \Gamma_*$, positive homogeneity gives $\mathfrak{m}(W_t)/\psi(W_t) \le \Gamma_*$, proving convergence. Compactness of the ψ-unit sphere and continuity of $\mathfrak{m}$ give the accumulation-point statement. □

B.2 Full-batch max–sum dominated norms

For a matrix A, write

$$\|A\|_{\max} := \max_{r,j} |A_{r,j}|, \qquad \|A\|_{\mathrm{sum}} := \sum_{r,j} |A_{r,j}|.$$

We call a matrix norm ψ *max–sum dominated* if

$$\|A\|_{\max} \le \psi(A) \le \|A\|_{\mathrm{sum}} \qquad \text{for every } A. \tag{14}$$

Let $R_1(X) := \max_{n\in[N]} \|x_n\|_1 < \infty$.

Lemma B.5 (Consequences of max–sum domination)**.** *Let ψ satisfy* (14)*, and let ψ^* be its dual norm. Then, for every matrix G, every h with $\|h\|_1 \le R_1(X)$, and every $A \in \mathbb{R}^{k\times d}$,*

$$\psi^*(G) \le \|G\|_{\mathrm{sum}}, \qquad \|Ah\|_\infty \le R_1(X)\psi(A).$$

Moreover, if $s \in \mathbb{R}^k$ is a probability vector and $H(s) := \mathrm{diag}(s) - ss^\top$, then for every $c \in [k]$,

$$h^\top A^\top H(s) Ah \le 4R_1(X)^2(1-s_c)\psi(A)^2. \tag{15}$$

Proof. Since the ψ-unit ball is contained in the entrywise max-norm unit ball,

$$\psi^*(G) = \max_{\psi(A)\le 1} \langle G, A\rangle \le \max_{\|A\|_{\max}\le 1} \langle G, A\rangle = \|G\|_{\mathrm{sum}}.$$

Also, $\|Ah\|_\infty \le \|A\|_{\max}\|h\|_1 \le R_1(X)\psi(A)$. For $v = Ah$, the softmax covariance bound gives $v^\top H(s)v \le 4(1-s_c)\|v\|_\infty^2$, which proves (15). □

Theorem B.6 (Full-batch max–sum dominated margin convergence)**.** *Let ψ be max–sum dominated. Under Assumption 2.5, normalized steepest descent*

$$\Delta_t \in \arg\max_{\psi(\Delta)\le 1} \langle \nabla L(W_t), \Delta\rangle, \qquad W_{t+1} = W_t - \eta_t \Delta_t$$

satisfies

$$\frac{\mathrm{mar}_X(W_t)}{\psi(W_t)} \longrightarrow \gamma_\psi(X).$$

Under Assumptions 2.5 and 2.6, the same conclusion holds for every fixed $\beta \in (0,1)$ under normalized momentum descent with $M_0 = \mathbf{0}$:

$$M_{t+1} = \beta M_t + (1-\beta)\nabla L(W_t), \qquad \Delta_t \in \arg\max_{\psi(\Delta)\le 1} \langle M_{t+1}, \Delta\rangle, \qquad W_{t+1} = W_t - \eta_t\Delta_t,$$

In either case, every accumulation point of $W_t/\psi(W_t)$ is a ψ-maximum-margin solution.

Proof. The geometry-dependent steps in the normalized steepest-descent and momentum proofs of Fan et al. (2025, Appendices D and E) use only norm duality, the lower bound $\psi^*(\nabla L(W)) \ge \gamma_\psi(X)\mathcal{G}(W)$, the upper bound $\psi^*(\nabla L(W)) \le 2R_1(X)\mathcal{G}(W)$, and the logit and Hessian controls in Lemma B.5. Substituting these bounds into their pre-rate scalar argument gives

$$\mathrm{mar}_X(W_t) \ge \gamma_\psi(X) S_t - r_t, \qquad \psi(W_t) \le \psi(W_0) + S_t,$$

where r_t is a finite-prefix term plus sums of the forms $\sum_s \eta_s^2$ and $\sum_s \eta_s \exp(-c\sum_{\tau<s}\eta_\tau)$; in the momentum case, Assumption 2.6 controls the additional geometric-memory term. Lemma B.3 gives $r_t = o(S_t)$, and Lemma B.4 with $\mathfrak{m} = \mathrm{mar}_X$ and $\Gamma = \Gamma_* = \gamma_\psi(X)$ proves the result. □

The full-batch Adam characterization used in Theorem 3.1 also extends to the learning-rate assumptions of the present paper.

Theorem B.7 (Full-batch Adam under the present learning-rate assumptions)**.** *Consider full-batch Adam without bias correction or a numerical-stability constant, initialized by $M_0 = V_0 = \mathbf{0}$:*

$$M_{t+1} = \beta_1 M_t + (1-\beta_1)\nabla L(W_t),$$
$$V_{t+1} = \beta_2 V_t + (1-\beta_2)\nabla L(W_t)^{\odot 2},$$
$$W_{t+1} = W_t - \eta_t M_{t+1} \oslash V_{t+1}^{\odot 1/2}.$$

Assume $0 < \beta_1 \le \beta_2 < 1$, Assumptions 2.5 and 2.6, and $\min_{c\in[k],\, j\in[d]} |\nabla L(W_0)[c,j]|^2 \ge \omega$ for some $\omega > 0$. Then, with

$$\gamma_{\max}(X) := \max_{\|W\|_{\max}\le 1} \mathrm{mar}_X(W),$$

one has

$$\frac{\mathrm{mar}_X(W_t)}{\|W_t\|_{\max}} \longrightarrow \gamma_{\max}(X), \tag{16}$$

and every accumulation point of $W_t/\|W_t\|_{\max}$ is a max-norm maximum-margin solution.

Proof. The pre-rate part of Fan et al. (2025, Appendix G) gives

$$\mathrm{mar}_X(W_t) \ge \gamma_{\max}(X) S_t - r_t, \qquad \|W_t\|_{\max} \le C_0 + S_t,$$

where r_t consists of a finite prefix, $\sum_s \eta_s^{3/2}$, and an exponentially weighted remainder. Lemma B.3 gives $r_t = o(S_t)$, and Lemma B.4 proves (16) and the accumulation-point conclusion. □

B.3 GEOMETRY-ADAPTED RANDOM-RESHUFFLING MOMENTUM

The full-batch max–sum argument above is sufficient for deterministic optimization, but using the same common feature radius $\max_n \|x_n\|_1$ in the stochastic analysis is unnecessarily coarse when comparing different geometries. We therefore work directly with the primal–dual pair (ψ, ψ^*) and define the atom radius

$$R_\psi(X) := \max_{n\in[N]} \max_{c\in[k]} \psi^*(e_c x_n^\top). \tag{17}$$

The stochastic results below hold for an arbitrary matrix norm ψ; max–sum domination is not required. Assume $b \mid N$, set $m := N/b$, and for every mini-batch $B \subset [N]$ of size b define $L_B(W) := b^{-1} \sum_{n\in B} \ell_n(W)$.

Lemma B.8 (Geometry-adapted cross-entropy bounds)**.** *For every matrix norm ψ, every sample x_n, every matrix A, and all classes $a, c \in [k]$,*

$$\|Ax_n\|_\infty \le R_\psi(X)\psi(A), \qquad \psi^*((e_a - e_c)x_n^\top) \le 2R_\psi(X). \tag{18}$$

For every W,

$$\gamma_\psi(X)\mathcal{G}(W) \le \psi^*(\nabla L(W)) \le 2R_\psi(X)\mathcal{G}(W). \tag{19}$$

For every sample n, every W, and every perturbation A,

$$D^2\ell_n(W)[A, A] \le 4R_\psi(X)^2\big(1 - \mathbb{S}(Wx_n)_{y_n}\big)\psi(A)^2. \tag{20}$$

Moreover, for all U, W,

$$|L(U) - L(W)| \le 2R_\psi(X)\psi(U - W). \tag{21}$$

Proof. For every $c \in [k]$, norm duality gives

$$|e_c^\top Ax_n| = |\langle A, e_c x_n^\top\rangle| \le \psi(A)\psi^*(e_c x_n^\top) \le R_\psi(X)\psi(A),$$

which proves the first bound in (18); the second follows from the triangle inequality. Writing $p_n(W) := \mathbb{S}(Wx_n)$, the sample gradient has the decomposition

$$\nabla\ell_n(W) = \sum_{c\neq y_n} p_{n,c}(W)(e_c - e_{y_n})x_n^\top,$$

so

$$\psi^*(\nabla\ell_n(W)) \le 2R_\psi(X)\big(1 - p_{n,y_n}(W)\big).$$

Averaging proves the upper bound in (19). For the lower bound, choose $A_\star$ with $\psi(A_\star) \le 1$ and margin $\gamma_\psi(X)$; then

$$-\nabla L(W) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\neq y_n} p_{n,c}(W)(e_{y_n} - e_c)x_n^\top$$

gives $\psi^*(\nabla L(W)) \ge \langle A_\star, -\nabla L(W)\rangle \ge \gamma_\psi(X)\mathcal{G}(W)$. For the Hessian bound, set $v := Ax_n$ and let C be categorical with law $p_n(W)$. Then

$$D^2\ell_n(W)[A, A] = \mathrm{Var}(v_C) \le \mathbb{E}[(v_C - v_{y_n})^2] \le 4(1 - p_{n,y_n}(W))\|v\|_\infty^2,$$

which proves (20). Finally, convexity and the upper gradient bound, together with $\mathcal{G} \le 1$, yield $L(U) - L(W) \le 2R_\psi(X)\psi(U - W)$; exchanging U, W proves (21). □

Lemma B.9 (Geometry-adapted softmax and gradient stability)**.** *For $U, W \in \mathbb{R}^{k\times d}$, define $h := R_\psi(X)\psi(U - W)$. Then, for every sample n and class c,*

$$e^{-2h} \le \frac{p_{n,c}(U)}{p_{n,c}(W)} \le e^{2h}, \qquad e^{-2h} \le \frac{1 - p_{n,y_n}(U)}{1 - p_{n,y_n}(W)} \le e^{2h}.$$

Consequently,

$$\mathcal{G}(U) \le e^{2R_\psi(X)\psi(U-W)}\mathcal{G}(W), \tag{22}$$

$$\psi^*(\nabla L(U) - \nabla L(W)) \le 2R_\psi(X)\Big(e^{2R_\psi(X)\psi(U-W)} - 1\Big)\mathcal{G}(W), \tag{23}$$

and, for every mini-batch B,

$$\psi^*(\nabla L_B(U) - \nabla L_B(W)) \le 2mR_\psi(X)\Big(e^{2R_\psi(X)\psi(U-W)} - 1\Big)\mathcal{G}(W). \tag{24}$$

Proof. Set $\delta_n := (U - W)x_n$. Lemma B.8 gives $\|\delta_n\|_\infty \le h$, and the softmax formula gives

$$\frac{p_{n,c}(U)}{p_{n,c}(W)} = e^{\delta_{n,c}} \frac{\sum_{a=1}^k e^{e_a^\top W x_n}}{\sum_{a=1}^k e^{e_a^\top W x_n + \delta_{n,a}}} \in [e^{-2h}, e^{2h}].$$

Summing over $c \neq y_n$ proves the second ratio and (22). If $p := p_n(W)$ and $q := p_n(U)$, the ratio bounds imply

$$\sum_{c \neq y_n} |q_c - p_c| \le (e^{2h} - 1)(1 - p_{y_n}), \qquad |q_{y_n} - p_{y_n}| \le (e^{2h} - 1)(1 - p_{y_n}).$$

Since $\nabla \ell_n(U) - \nabla \ell_n(W) = (q - p)x_n^\top$,

$$\psi^*((q - p)x_n^\top) \le \sum_{c=1}^k |q_c - p_c| \psi^*(e_c x_n^\top) \le 2R_\psi(X)(e^{2h} - 1)(1 - p_{y_n}).$$

Averaging over all samples proves (23); averaging over B, enlarging the resulting nonnegative sample sum to $[N]$, and using $N/b = m$ proves (24). □

Lemma B.10 (Random-reshuffling noise and cancellation)**.** *For a mini-batch B of size b, define $\Xi_B(W) := \nabla L_B(W) - \nabla L(W)$. Then*

$$\psi^*(\Xi_B(W)) \le 2(m - 1)R_\psi(X)\mathcal{G}(W).$$

Moreover, let $\{B_t\}_{t \ge 0}$ be any random-reshuffling sequence, so that $B_{qm}, \dots, B_{qm+m-1}$ partition $[N]$ for every epoch q, and define

$$E_t(W) := (1 - \beta) \sum_{\tau=0}^t \beta^\tau \big(\nabla L_{B_{t-\tau}}(W) - \nabla L(W)\big).$$

Then, deterministically for every reshuffling realization,

$$\psi^*(E_t(W)) \le (1 - \beta)m(m^2 - 1)R_\psi(X)\mathcal{G}(W).$$

Proof. When $m \ge 2$, let $B^c = [N] \setminus B$. The finite-population identity

$$\Xi_B(W) = \frac{m - 1}{m}\big(\nabla L_B(W) - \nabla L_{B^c}(W)\big)$$

and the single-sample bound in the proof of Lemma B.8 give

$$\psi^*(\Xi_B(W)) \le \frac{m - 1}{N} \sum_{n=1}^N \psi^*(\nabla \ell_n(W)) \le 2(m - 1)R_\psi(X)\mathcal{G}(W),$$

while $m = 1$ is immediate.

For the second assertion, write $\Xi_t(W) := \nabla L_{B_t}(W) - \nabla L(W)$. Every complete epoch satisfies $\sum_{r=0}^{m-1} \Xi_{qm+r}(W) = \mathbf{0}$. Put $K_\Xi(W) := 2(m - 1)R_\psi(X)\mathcal{G}(W)$. For a complete epoch contained in $\{0, \dots, t\}$, subtracting the constant coefficient β^{t-qm} and using $1 - \beta^r \le r(1 - \beta)$ gives

$$\psi^*\left(\sum_{r=0}^{m-1} \beta^{t-qm-r} \Xi_{qm+r}(W)\right) \le K_\Xi(W)\beta^{t-(q+1)m}(1 - \beta)\frac{m(m - 1)}{2}.$$

Summing over complete epochs and using $(1 - \beta)/(1 - \beta^m) \le 1$, then adding the at most m terms in the current incomplete epoch, gives

$$\psi^*(E_t(W)) \le (1 - \beta)\frac{m(m + 1)}{2} K_\Xi(W) = (1 - \beta)m(m^2 - 1)R_\psi(X)\mathcal{G}(W).$$

□

Consider now the moving-weight random-reshuffling momentum process

$$M_{t+1} = \beta M_t + (1-\beta)\nabla L_{B_t}(W_t), \qquad \Delta_t \in \arg\max_{\psi(\Delta)\le 1}\langle M_{t+1}, \Delta\rangle, \qquad W_{t+1} = W_t - \eta_t \Delta_t,$$

with $M_0 = \mathbf{0}$. Since $\psi(\Delta_t) \le 1$,

$$\psi(W_t - W_{t-\tau}) \le \sum_{q=1}^{\tau} \eta_{t-q}. \tag{25}$$

Proposition B.11 (Geometry-adapted random-reshuffling momentum control)**.** *Assume Assumptions 2.5 and 2.6. For every fixed finite dataset, fixed matrix norm ψ, and fixed $\beta \in (0,1)$, there exist $t_{0,\psi,\beta} < \infty$ and $c_{\psi,\beta} < \infty$ such that, for all $t \ge t_{0,\psi,\beta}$,*

$$\psi^*(M_{t+1} - \nabla L(W_t)) \le \Big[2m(1-\beta)R_\psi(X)c_{\psi,\beta}\eta_t + (1-\beta)m(m^2-1)R_\psi(X) + 2R_\psi(X)\beta^{t+1}\Big]\mathcal{G}(W_t). \tag{26}$$

Define

$$\rho_{\psi,\beta,b}(X) := \gamma_\psi(X) - 2(1-\beta)m(m^2-1)R_\psi(X), \tag{27}$$

$$C_{\psi,\beta,b} := 4m(1-\beta)R_\psi(X)c_{\psi,\beta} + 2R_\psi(X)^2 e^{2R_\psi(X)\eta_0}, \qquad D_\psi := 4R_\psi(X).$$

Then, for all sufficiently large t,

$$L(W_{t+1}) \le L(W_t) - \eta_t\Big(\rho_{\psi,\beta,b}(X) - C_{\psi,\beta,b}\eta_t - D_\psi\beta^{t/2}\Big)\mathcal{G}(W_t). \tag{28}$$

Proof. Unrolling the momentum recursion and adding and subtracting frozen batch gradients at W_t gives

$$\begin{aligned} M_{t+1} - \nabla L(W_t) =& (1-\beta)\sum_{\tau=0}^{t}\beta^\tau\Big(\nabla L_{B_{t-\tau}}(W_{t-\tau}) - \nabla L_{B_{t-\tau}}(W_t)\Big) \\ &+ (1-\beta)\sum_{\tau=0}^{t}\beta^\tau\Big(\nabla L_{B_{t-\tau}}(W_t) - \nabla L(W_t)\Big) - \beta^{t+1}\nabla L(W_t). \end{aligned}$$

Lemma B.9 and (25) bound the first term by

$$2m(1-\beta)R_\psi(X)\mathcal{G}(W_t)\sum_{\tau=0}^{t}\beta^\tau\left(\exp\left(2R_\psi(X)\sum_{q=1}^{\tau}\eta_{t-q}\right) - 1\right).$$

Assumption 2.6, applied with $c_1 = 2R_\psi(X)$, bounds the final sum by $c_{\psi,\beta}\eta_t$ for all sufficiently large t. Lemma B.10 controls the second term, and the upper gradient bound in (19) controls the last, proving (26).

For the descent recurrence, the steepest-descent property and norm duality give

$$\langle \nabla L(W_t), \Delta_t\rangle \ge \psi^*(\nabla L(W_t)) - 2\psi^*(M_{t+1} - \nabla L(W_t)). \tag{29}$$

The mean-value Taylor formula, (20), and $\psi(\Delta_t) \le 1$ bound the quadratic remainder in $L(W_t - \eta_t\Delta_t)$ by

$$2\eta_t^2 R_\psi(X)^2 \sup_{0\le s\le 1}\mathcal{G}(W_t - s\eta_t\Delta_t).$$

Lemma B.9 bounds the supremum by $e^{2R_\psi(X)\eta_0}\mathcal{G}(W_t)$. Substituting (19) and (26) into (29), adding the quadratic remainder, and using $\beta^{t+1} \le \beta^{t/2}$ proves (28). □

Corollary B.12 (Geometry-adapted random-reshuffling momentum margin)**.** *Under Assumptions 2.5 and 2.6, suppose $\rho_{\psi,\beta,b}(X) > 0$. Then*

$$\liminf_{t\to\infty}\frac{\mathrm{mar}_X(W_t)}{\psi(W_t)} \ge \rho_{\psi,\beta,b}(X). \tag{30}$$

Consequently, with

$$\varepsilon_{\psi,\beta,b}(X) := 2(1-\beta)m(m^2-1)R_\psi(X),$$

one has

$$d_{\cos}\Big(W_t, \mathcal{U}^X_{\psi,\varepsilon_{\psi,\beta,b}(X)}\Big) \longrightarrow 0. \tag{31}$$

More quantitatively, there exist a finite t_2 and $C_{\mathrm{ga}} < \infty$ such that

$$\rho_{\psi,\beta,b}(X) - \frac{\mathrm{mar}_X(W_t)}{\psi(W_t)} \le C_{\mathrm{ga}} \frac{\mathcal{R}_t^{\psi,\beta,b}}{S_t}, \tag{32}$$

where

$$\begin{aligned}\mathcal{R}_t^{\psi,\beta,b} :=& 1 + \sum_{s=0}^{t_2-1} \eta_s + C_{\psi,\beta,b} \sum_{s=t_2}^{t-1} \eta_s^2 + \frac{D_\psi \eta_0}{1-\sqrt{\beta}} \\ &+ \sum_{s=t_2}^{t-1} \eta_s \exp\left(-\frac{\rho_{\psi,\beta,b}(X)}{4} \sum_{\tau=t_2}^{s-1} \eta_\tau\right).\end{aligned}$$

Proof. Write $\rho := \rho_{\psi,\beta,b}(X) > 0$. Since $\eta_t \to 0$ and $\beta^{t/2} \to 0$, Proposition B.11 gives a finite t_1 after which the coefficient of $\eta_t \mathcal{G}(W_t)$ is at least $\rho/2$. The loss Lipschitz bound (21), the divergence $S_t \to \infty$, and the standard low-loss entrance argument give a finite $t_2 \ge t_1$ after which all empirical margins are nonnegative. The usual cross-entropy loss-to-margin conversion then yields

$$\mathrm{mar}_X(W_t) \ge \rho \sum_{s=t_2}^{t-1} \eta_s \frac{\mathcal{G}(W_s)}{L(W_s)} - C_{\psi,\beta,b} \sum_{s=t_2}^{t-1} \eta_s^2 - \frac{D_\psi \eta_0}{1-\sqrt{\beta}} - C_0,$$

where $C_0 < \infty$ absorbs the finite prefix. The post-entrance descent recurrence and the standard cross-entropy proxy inequalities give

$$\frac{\mathcal{G}(W_s)}{L(W_s)} \ge 1 - \exp\left(-\frac{\rho}{4} \sum_{\tau=t_2}^{s-1} \eta_\tau\right).$$

Hence $\mathrm{mar}_X(W_t) \ge \rho S_t - C_1 \mathcal{R}_t^{\psi,\beta,b}$ for some $C_1 < \infty$. Lemma B.3 gives $\mathcal{R}_t^{\psi,\beta,b} = o(S_t)$, while $\psi(W_t) \le \psi(W_0) + S_t$ follows directly from the normalized update. Lemma B.4 proves (32) and (30), and Lemma B.1 gives (31). □

For the three geometries used in the main text, the atom radius is explicit:

$$R_{\mathrm{row}}(X) = \max_n \|x_n\|_2, \qquad R_{\mathrm{sp}}(X) = \max_n \|x_n\|_2, \qquad R_{\max}(X) = \max_n \|x_n\|_1. \tag{33}$$

Indeed, the duals of the row, spectral, and entrywise max norms are the sum-of-row-Euclidean, nuclear, and entrywise sum norms, respectively, while $e_c x_n^\top$ has one nonzero row and rank one.

B.4 Specialization to row normalization

Define

$$\|A\|_{\mathrm{row}} := \max_{r \in [k]} \|A_{r,:}\|_2.$$

For a matrix G, define its row-wise normalization by

$$\big(N_{\mathrm{row}}(G)\big)_{r,:} := \begin{cases} G_{r,:}/\|G_{r,:}\|_2, & G_{r,:} \neq \mathbf{0}, \\ \mathbf{0}, & G_{r,:} = \mathbf{0}. \end{cases}$$

Proposition B.13 (Row normalization is row-norm steepest descent)**.** *For every $A, G \in \mathbb{R}^{k\times d}$,*

$$\|A\|_{\max} \le \|A\|_{\mathrm{row}} \le \|A\|_{\mathrm{sum}}, \tag{34}$$

and

$$N_{\mathrm{row}}(G) \in \arg\max_{\|\Delta\|_{\mathrm{row}} \le 1} \langle G, \Delta \rangle, \qquad \max_{\|\Delta\|_{\mathrm{row}} \le 1} \langle G, \Delta \rangle = \sum_{r=1}^{k} \|G_{r,:}\|_2.$$

Thus row-wise normalized gradient methods are normalized steepest-descent methods for the row norm.

Proof. Every entry is bounded by the Euclidean norm of its row, while $\max_r \|A_{r,:}\|_2 \le \sum_r \|A_{r,:}\|_1 = \|A\|_{\mathrm{sum}}$, proving (34). The constraint $\|\Delta\|_{\mathrm{row}} \le 1$ is equivalent to $\|\Delta_{r,:}\|_2 \le 1$ for every row r, so

$$\max_{\|\Delta\|_{\mathrm{row}} \le 1} \langle G, \Delta \rangle = \sum_{r=1}^{k} \max_{\|u\|_2 \le 1} \langle G_{r,:}, u \rangle = \sum_{r=1}^{k} \|G_{r,:}\|_2,$$

and $N_{\mathrm{row}}(G)$ attains the maximum. □

Proposition B.14 (Row-normalized full-batch margin convergence)**.** *Consider any fixed linearly separable multiclass dataset with bounded features. Under Assumption 2.5, row-wise normalized GD satisfies*

$$\frac{\mathrm{mar}_X(W_t)}{\|W_t\|_{\mathrm{row}}} \longrightarrow \gamma_{\mathrm{row}}(X), \qquad \inf_{V \in \mathcal{U}^X_{\mathrm{row}}} \left\| \frac{W_t}{\|W_t\|_{\mathrm{F}}} - \frac{V}{\|V\|_{\mathrm{F}}} \right\|_{\mathrm{F}} \longrightarrow 0.$$

If Assumption 2.6 also holds, the same two conclusions hold for row-wise normalized GD with every fixed momentum parameter $\beta \in (0, 1)$.

Proof. Proposition B.13 shows that the row norm is max–sum dominated and that row normalization is its exact steepest direction. Theorem B.6 therefore gives normalized-margin convergence, and Lemma B.2 gives $d_{\cos}(W_t, \mathcal{U}^X_{\mathrm{row}}) \to 0$. Since

$$\inf_{V \in \mathcal{U}^X_{\mathrm{row}}} \left\| \frac{W_t}{\|W_t\|_{\mathrm{F}}} - \frac{V}{\|V\|_{\mathrm{F}}} \right\|^2_{\mathrm{F}} = 2 d_{\cos}(W_t, \mathcal{U}^X_{\mathrm{row}}),$$

the Frobenius direction-set convergence follows. □

The stochastic counterpart follows from the geometry-adapted random-reshuffling result above and is stated separately because the approximation gap depends on the batch size and momentum.

Proposition B.15 (Random-reshuffling row-wise normalized SGDM margin convergence)**.** *Consider any fixed linearly separable multiclass dataset with bounded features. Run random-reshuffling row-wise normalized SGDM with $b \mid N$, $m := N/b$, and fixed $\beta \in (0, 1)$. Define*

$$R_{\mathrm{row}}(X) := \max_{n \in [N]} \|x_n\|_2, \qquad \varepsilon_{\mathrm{row},\beta,b}(X) := 2(1-\beta) m (m^2 - 1) R_{\mathrm{row}}(X).$$

Under Assumptions 2.5 and 2.6, if $\varepsilon_{\mathrm{row},\beta,b}(X) < \gamma_{\mathrm{row}}(X)$, then

$$\liminf_{t \to \infty} \frac{\mathrm{mar}_X(W_t)}{\|W_t\|_{\mathrm{row}}} \ge \gamma_{\mathrm{row}}(X) - \varepsilon_{\mathrm{row},\beta,b}(X), \tag{35}$$

and

$$d_{\cos}\Big(W_t, \mathcal{U}^X_{\mathrm{row},\varepsilon_{\mathrm{row},\beta,b}(X)}\Big) \longrightarrow 0. \tag{36}$$

Proof. Apply Corollary B.12 with $\psi = \|\cdot\|_{\mathrm{row}}$. Equation (33) gives $R_\psi(X) = R_{\mathrm{row}}(X)$, so (27) becomes $\gamma_{\mathrm{row}}(X) - \varepsilon_{\mathrm{row},\beta,b}(X)$, yielding (35). Equation (36) follows from (31). □

C Proofs for Deterministic Optimizers

This appendix proves Theorem 3.1 from optimizer-selected margin geometry to population accuracy. We first establish the all-aspect-ratio Row-versus-Spectral and Row-versus-Max geometry: Row representatives align uniformly with the population mean differences, Max representatives have a nonvanishing clipped-coordinate angular gap, and Spectral representatives are controlled through the reduced polar factor of the centered empirical mean matrix, with hard-edge truncation at $\alpha = 1$. We then convert these geometric statements into ordinary multiclass population-error bounds and prove the optimizer-level accuracy ordering. We also derive deterministic population-error exponents that quantify the separation between the three geometries.

For any matrix norm ψ, define the unit-margin feasible set and the corresponding minimum-norm representative set by

$$\mathcal{F}_X := \left\{W \in \mathbb{R}^{k\times d} : \operatorname{mar}_X(W) \geq 1\right\}, \qquad \mathcal{M}_\psi^X := \operatorname*{arg\,min}_{W\in\mathcal{F}_X} \psi(W).$$

Whenever X is linearly separable, $\mathcal{F}_X$ is nonempty, the minimum is attained, and we define

$$r_\psi^*(X) := \min_{W\in\mathcal{F}_X} \psi(W) = \frac{1}{\gamma_\psi(X)}, \qquad \mathcal{U}_\psi^X = \left\{aW : a > 0,\ W \in \mathcal{M}_\psi^X\right\}. \tag{37}$$

For $q \in \{\mathrm{row}, \mathrm{sp}, \max\}$, we abbreviate $\mathcal{M}_q^X := \mathcal{M}_{\psi_q}^X$ and $r_q^*(X) := r_{\psi_q}^*(X)$. For every feasible classifier and every ordered pair $i \neq j$, define

$$u_{ij}^W := \frac{w_i - w_j}{\|w_i - w_j\|_2}, \qquad u_{ij}^\mu := \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|_2}, \qquad D_{\text{avg-bdry}}(W) := \frac{1}{k(k-1)} \sum_{i\neq j} \|u_{ij}^W - u_{ij}^\mu\|_2.$$

These quantities are invariant under positive rescaling of W and under common row shifts.

Lemma C.1 (Exact invariance under common feature rescaling)**.** *Let $c > 0$, let $cX := \{(cx_n, y_n)\}_{n=1}^N$, and let ψ be any matrix norm. Then*

$$\mathcal{F}_{cX} = c^{-1}\mathcal{F}_X, \qquad \mathcal{M}_\psi^{cX} = c^{-1}\mathcal{M}_\psi^X, \qquad \mathcal{U}_\psi^{cX} = \mathcal{U}_\psi^X.$$

The normalized pairwise row directions and $D_{\text{avg-bdry}}$ are unchanged. If the test inputs are also multiplied by c, then $\mathrm{Acc}_{\pi,d}(c^{-1}W; cX^{\mathrm{te}}) = \mathrm{Acc}_{\pi,d}(W; X^{\mathrm{te}})$.

Proof. The constraint $(w_{y_n} - w_j)^\top (cx_n) \geq 1$ is equivalent, after writing $W = c^{-1}V$, to $(v_{y_n} - v_j)^\top x_n \geq 1$. Positive homogeneity of ψ proves the identities for the feasible and minimum-norm sets, and the ray identity follows from (37). Finally, $(c^{-1}w_j)^\top (cx) = w_j^\top x$, so all predictions agree. □

Apply Lemma C.1 with $c = d^{-1/2}$. Thus, throughout the geometric and accuracy appendices, we work with

$$\widetilde{\mu}_i := \frac{\mu_i}{\sqrt{d}}, \qquad \widetilde{z}_{i,s} := \frac{z_{i,s}}{\sqrt{d}}, \qquad \widetilde{x}_{i,s} := \frac{x_{i,s}}{\sqrt{d}}, \tag{38}$$

and suppress tildes. Under this convention, $\mu_i, z_{i,s} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d/d)$ and $X^{\mathrm{te}} \mid (Y = i) \sim \mathcal{N}(\mu_i, r_{\mathrm{te}}^2 I_d/d)$. This is only a proof normalization: the direction sets, accuracy values, and radius-to-margin ratios used in the optimizer theorems are identical to those in the unnormalized model of Assumption 2.1.

Let $G \sim \mathcal{N}(0,1)$, and define

$$A_{\max}(\tau) := \mathbb{E}[|G| \min\{\tau|G|, 1\}], \qquad B_{\max}(\tau) := \mathbb{E}\big[\min\{\tau^2 G^2, 1\}\big]. \tag{39}$$

Let $\tau_* \in (0,1)$ be the unique solution of $A_{\max}(\tau_*) = 1/\sqrt{\pi}$, and define

$$C_{\max} := \big(2\pi B_{\max}(\tau_*)\big)^{-1/2}, \qquad \kappa_{\max} := \sqrt{2}\, C_{\max} < 1, \qquad \delta_{\max} := \sqrt{2 - 2\kappa_{\max}} > 0. \tag{40}$$

For $\alpha > 0$, let ν_α be the mean-one Marchenko–Pastur law with parameter $1/\alpha$, including its zero atom when $\alpha < 1$:

$$\nu_\alpha = (1-\alpha)_+ \delta_0 + \frac{\alpha}{2\pi x} \sqrt{(b_\alpha - x)(x - a_\alpha)}\, \mathbf{1}_{[a_\alpha, b_\alpha]}(x)\, dx,$$
$$a_\alpha = (1 - \alpha^{-1/2})^2, \qquad b_\alpha = (1 + \alpha^{-1/2})^2.$$

Set

$$\rho_\alpha := \alpha \wedge 1, \qquad m_{1/2}(\alpha) := \int \sqrt{x}\, d\nu_\alpha(x), \qquad s_\alpha := \frac{m_{1/2}(\alpha)}{\sqrt{\alpha}}, \qquad \kappa_{\mathrm{sp}}(\alpha) := \frac{m_{1/2}(\alpha)}{\sqrt{\rho_\alpha}}, \tag{41}$$

and

$$\delta_{\mathrm{sp}}(\alpha) := \sqrt{2 - 2\kappa_{\mathrm{sp}}(\alpha)} > 0.$$

Strict positivity follows from strict Cauchy–Schwarz on the positive part of ν_α: $m_{1/2}(\alpha)^2 < \rho_\alpha \int x\, d\nu_\alpha(x) = \rho_\alpha$. At $\alpha = 1$, $m_{1/2}(1) = 8/(3\pi)$, so $\delta_{\mathrm{sp}}(1) = \sqrt{2 - 16/(3\pi)}$.

Theorem C.2 (All-aspect deterministic margin geometry)**.** *Under Assumptions 2.1 and 2.2, there are deterministic $\omega_d \downarrow 0$ and events $\mathcal{G}_d^{\mathrm{det}}$ such that, for every sufficiently large d,*

$$\mathbb{P}(\mathcal{G}_d^{\mathrm{det}}) \geq 1 - e^{-\sqrt{d}},$$

and, on $\mathcal{G}_d^{\mathrm{det}}$, $\mathcal{F}_X \neq \varnothing$ and

$$\sup_{U \in \mathcal{U}_{\mathrm{row}}^X} \max_{i \neq j} \|u_{ij}^U - u_{ij}^\mu\|_2 \leq \omega_d, \tag{42}$$

$$\sup_{U \in \mathcal{U}_{\mathrm{row}}^X} D_{\text{avg-bdry}}(U) \leq \omega_d, \tag{43}$$

$$\inf_{U \in \mathcal{U}_{\mathrm{sp}}^X} D_{\text{avg-bdry}}(U) \geq \delta_{\mathrm{sp}}(\alpha) - \omega_d, \tag{44}$$

$$\inf_{U \in \mathcal{U}_{\max}^X} D_{\text{avg-bdry}}(U) \geq \delta_{\max} - \omega_d. \tag{45}$$

Consequently, for every sufficiently large d on $\mathcal{G}_d^{\mathrm{det}}$,

$$\sup_{U \in \mathcal{U}_{\mathrm{row}}^X} D_{\text{avg-bdry}}(U) < \min\left\{ \inf_{V \in \mathcal{U}_{\mathrm{sp}}^X} D_{\text{avg-bdry}}(V), \inf_{V \in \mathcal{U}_{\max}^X} D_{\text{avg-bdry}}(V) \right\}. \tag{46}$$

C.1 Reduction from Direction Sets to Minimum-norm Representatives

For every feasible matrix, all pairwise row differences are nonzero. Indeed, if $w_i = w_j$ for some $i \neq j$, then the constraint $(w_i - w_j)^\top x_{i,s} \geq 1$ fails. Thus $D_{\text{avg-bdry}}$ is well defined on $\mathcal{F}_X$ and on all three minimum-norm solution sets.

Lemma C.3 (Scale-invariant reduction)**.** *For every $\psi \in \{\psi_{\mathrm{row}}, \psi_{\mathrm{sp}}, \psi_{\max}\}$,*

$$\sup_{U \in \mathcal{U}_\psi^X} D_{\text{avg-bdry}}(U) = \sup_{W \in \mathcal{M}_\psi^X} D_{\text{avg-bdry}}(W),$$

$$\inf_{U \in \mathcal{U}_\psi^X} D_{\text{avg-bdry}}(U) = \inf_{W \in \mathcal{M}_\psi^X} D_{\text{avg-bdry}}(W).$$

Proof. By (37), $\mathcal{U}_\psi^X = \{aW : a > 0,\ W \in \mathcal{M}_\psi^X\}$. The conclusion follows from the positive scale invariance of $D_{\text{avg-bdry}}$. □

We will therefore work with the minimum-norm representatives $\mathcal{M}_{\mathrm{row}}^X$, $\mathcal{M}_{\mathrm{sp}}^X$, and $\mathcal{M}_{\max}^X$, and invoke Lemma C.3 only at the end. The max-norm constants are those in (40).

C.2 A Quantitative Common Gaussian-cloud Event

For each class, define

$$\bar{z}_i := \frac{1}{n_i} \sum_{s=1}^{n_i} z_{i,s}, \qquad m_i := \frac{1}{n_i} \sum_{s=1}^{n_i} x_{i,s} = \mu_i + r\bar{z}_i, \qquad e_{i,s} := x_{i,s} - m_i = r(z_{i,s} - \bar{z}_i). \tag{47}$$

Then $x_{i,s} = m_i + e_{i,s}$, $\sum_s e_{i,s} = \mathbf{0}$, and m_i is independent of the centered residual family $(e_{i,1}, \ldots, e_{i,n_i})$. Moreover,

$$\mathrm{Cov}(e_{i,s}) = r^2 \left(1 - \frac{1}{n_i}\right) \frac{I_d}{d}.$$

For $i \neq j$, write $u_{ij}^m := (m_i - m_j)/\|m_i - m_j\|_2$. For $u \in \mathbb{S}^{d-1}$, set

$$H_u(\tau) := \frac{1}{d} \sum_{\ell=1}^{d} \sqrt{d}\,|u_\ell| \min\{\tau\sqrt{d}\,|u_\ell|, 1\}, \qquad J_u(\tau) := \frac{1}{d} \sum_{\ell=1}^{d} \min\{\tau^2 d u_\ell^2, 1\}. \tag{48}$$

Proposition C.4 (Stretched-exponential common input)**.** *Under Assumptions 2.1 and 2.2, there are deterministic $\varepsilon_d \downarrow 0$, universal $C < \infty$, a constant $c = c(r_0) > 0$, and events $\mathcal{E}_d^{\mathrm{det}}$ such that*

$$\mathbb{P}\big((\mathcal{E}_d^{\mathrm{det}})^c\big) \leq C \exp\Big[-c\sqrt{d \log(2N(d)(d+1))}\Big] \tag{49}$$

and the following statements hold simultaneously on $\mathcal{E}_d^{\mathrm{det}}$:

(H1) $\max_i \|m_i - \mu_i\|_2 \le \varepsilon_d$.

(H2) $\max_i |\|m_i\|_2 - 1| \le \varepsilon_d$ *and* $\max_{i\neq j} |m_i^\top m_j| \le \varepsilon_d$.

(H3) $\max_{i,j,s} |m_j^\top e_{i,s}| \le \varepsilon_d$.

(H4) $\max_{i\neq j} \|u_{ij}^m - u_{ij}^\mu\|_2 \le \varepsilon_d$.

(H5) The max-norm optimum is finite and satisfies

$$2\sqrt{d}\, r^*_{\max}(X) \le \sqrt{2\pi} + \varepsilon_d.$$

(H6) Uniformly over ordered pairs,

$$\sup_{0\le\tau\le1} |H_{u_{ij}^m}(\tau) - A_{\max}(\tau)| + \sup_{0\le\tau\le1} |J_{u_{ij}^m}(\tau) - B_{\max}(\tau)| \le \varepsilon_d. \tag{50}$$

Proof. Put $L_d := \log(2N(d)(d+1))$ and $a_d := (L_d/d)^{1/4}$. The assumptions $\log N = o(d)$ and $n_{\min} \to \infty$ imply $a_d \to 0$. All normalized Gaussian vectors below obey the shell bound

$$\mathbb{P}\left(\left|\frac{\|g\|_2}{\sqrt{d}} - 1\right| > t\right) \le 2e^{-cdt^2}, \qquad 0 < t < 1, \tag{51}$$

and every centered Gaussian scalar with variance at most C/d obeys $\mathbb{P}(|Z| > t) \le 2e^{-cdt^2/C}$.

Since $\bar{z}_i \overset{\mathrm{d}}{=} g_i/\sqrt{n_i d}$, the event $\max_i \|\bar{z}_i\|_2 \le 2/\sqrt{n_{\min}}$ fails with probability at most $2Ne^{-cd}$. Thus

$$\max_i \|m_i - \mu_i\|_2 \le \frac{2r_0}{\sqrt{n_{\min}}} = o(1) \tag{52}$$

on an event whose failure probability is smaller than the right-hand side of (49). Applying (51) to $\mu_i \sim \mathcal{N}(\mathbf{0}, I_d/d)$, with threshold Ca_d, and taking a union bound proves $\max_i |\|\mu_i\|_2 - 1| \le Ca_d$. Conditionally on μ_i, the variable $\mu_i^\top \mu_j$ is centered Gaussian with variance $\|\mu_i\|_2^2/d$. Another union bound over at most N^2 ordered pairs gives $\max_{i\neq j} |\mu_i^\top \mu_j| \le Ca_d$. Combining these estimates with (52) proves (H1) and (H2).

Conditionally on all empirical means, $m_j^\top e_{i,s}$ is centered Gaussian with variance at most $r_0^2\|m_j\|_2^2/d$. On the event $\max_j \|m_j\|_2 \le 2$, a union bound over at most N^2 triples gives $\max_{i,j,s} |m_j^\top e_{i,s}| \le Ca_d$ with the required failure probability. This proves (H3). Items (H1)–(H2) imply $\inf_{i\neq j} \|m_i - m_j\|_2 \ge 1$ and $\inf_{i\neq j} \|\mu_i - \mu_j\|_2 \ge 1$ for all sufficiently large d. The normalization map is Lipschitz on vectors with norm at least one, so (H1) gives (H4).

For (H5), let $s_i := \operatorname{sign}(m_i) \in \{-1, 1\}^d$. Because each m_i is isotropic Gaussian, s_i is a uniform sign vector, and it is independent of every residual in its own class. Bernstein's inequality for $\|m_i\|_1$, Gaussian tails for $s_j^\top m_i$ when $i \neq j$, conditional Gaussian tails for $s_j^\top e_{i,s}$, and a union bound give

$$\min_i \|m_i\|_1 \ge \left(\sqrt{\frac{2}{\pi}} - o(1)\right)\sqrt{d}, \qquad \max_{i\neq j} |s_j^\top m_i| + \max_{i,j,s} |s_j^\top e_{i,s}| = o(\sqrt{d}) \tag{53}$$

outside an event bounded as in (49). The matrix with rows $w_i = t_d s_i$, where t_d is the reciprocal of the minimum bracket in

$$(s_i - s_j)^\top x_{i,s} \ge \|m_i\|_1 - |s_j^\top m_i| - 2\max_\ell |s_\ell^\top e_{i,s}|,$$

is feasible. Equation (53) gives $t_d \le (\sqrt{\pi/2} + o(1))/\sqrt{d}$, which is exactly (H5).

For (H6), each difference $m_i - m_j$ is a centered isotropic Gaussian vector, even when $n_i \neq n_j$, so u_{ij}^m is exactly uniform on $\mathbb{S}^{d-1}$. For one uniform direction, the two coordinate empirical processes in (48) concentrate about $A_{\max}$ and $B_{\max}$ at rate $e^{-cd\eta^2}$ for any fixed τ. On the event $d^{-1}\sum_\ell du_\ell^2 = 1$, both processes are uniformly Lipschitz in τ. A mesh of $[0, 1]$ with spacing a_d, followed by a union bound over the mesh and at most N^2 ordered pairs, yields (50) with error Ca_d and failure probability bounded by (49). Taking ε_d to dominate Ca_d, $2r_0/\sqrt{n_{\min}}$, and the deterministic vanishing envelopes in (53) completes the proof. □

Lemma C.5 (Normalization stability). *Let $x, y \in \mathbb{R}^d$ with $y \neq \mathbf{0}$. If $\|x - y\|_2 \leq \|y\|_2/2$, then*

$$\left\| \frac{x}{\|x\|_2} - \frac{y}{\|y\|_2} \right\|_2 \leq \frac{4\|x - y\|_2}{\|y\|_2}.$$

Proof. The reverse triangle inequality gives $\|x\|_2 \geq \|y\|_2/2$. Adding and subtracting $y/\|x\|_2$ and applying the reverse triangle inequality once more gives the displayed bound. □

C.3 Row-norm Representatives Align with the Population Mean Differences

The row-norm part is deterministic once the empirical means and residuals satisfy the small-error conditions in Proposition C.4.

Proposition C.6 (Deterministic row-norm rigidity). *Fix deterministic samples $x_{i,s} = m_i + e_{i,s}$ with $\sum_s e_{i,s} = \mathbf{0}$ for every i. Suppose that, for some $0 < h < 1/20$,*

$$\max_i \big| \|m_i\|_2 - 1 \big| \leq h, \qquad \max_{i \neq j} |m_i^\top m_j| \leq h, \qquad \max_{i,j,s} |m_j^\top e_{i,s}| \leq h.$$

Then $\mathcal{F}_X$ is nonempty, and every $W \in \mathcal{M}_{\mathrm{row}}^X$ satisfies

$$\max_i \|w_i - m_i\|_2 \leq C \left(h + \frac{1}{k} \right)^{1/4} \tag{54}$$

for a universal constant C.

Proof. For $i \neq j$ and $s \leq n_i$,

$$\begin{aligned}(m_i - m_j)^\top x_{i,s} &= \|m_i\|_2^2 - m_j^\top m_i + (m_i - m_j)^\top e_{i,s} \\ &\geq (1 - h)^2 - h - 2h \geq 1 - 5h.\end{aligned}$$

Hence the matrix with rows $\widetilde{w}_i = m_i/(1 - 5h)$ is feasible. If $R^* := \max_i \|w_i\|_2$ is the optimal row radius, then

$$R^* \leq \frac{1 + h}{1 - 5h} \leq 1 + Ch. \tag{55}$$

Let $\bar{w} := k^{-1} \sum_{r=1}^k w_r$. For any feasible W, averaging the constraints over $s = 1, \ldots, n_i$ and summing over $j \neq i$ gives

$$\sum_{j \neq i} (w_i - w_j)^\top m_i \geq k - 1.$$

Since $\sum_{j \neq i} w_j = k\bar{w} - w_i$, this is equivalent to

$$(w_i - \bar{w})^\top m_i \geq 1 - \frac{1}{k}. \tag{56}$$

Consequently,

$$\|w_i - \bar{w}\|_2 \geq L := \frac{1 - 1/k}{1 + h}.$$

For a row-norm minimizer, the variance identity yields

$$L^2 \leq \frac{1}{k} \sum_{i=1}^k \|w_i - \bar{w}\|_2^2 = \frac{1}{k} \sum_{i=1}^k \|w_i\|_2^2 - \|\bar{w}\|_2^2 \leq (R^*)^2 - \|\bar{w}\|_2^2.$$

Combining this with (55) gives

$$\|\bar{w}\|_2^2 \leq C \left(h + \frac{1}{k} \right).$$

Using (56) once more,

$$w_i^\top m_i \geq 1 - \frac{1}{k} + \bar{w}^\top m_i \geq 1 - C \left(h + \frac{1}{k} \right)^{1/2}.$$

Together with $\|w_i\|_2 \le 1 + Ch$ and $\|m_i\|_2 \le 1 + h$, this implies

$$\|w_i - m_i\|_2^2 = \|w_i\|_2^2 + \|m_i\|_2^2 - 2w_i^\top m_i \le C\left(h + \frac{1}{k}\right)^{1/2}.$$

Taking square roots proves (54). □

Proposition C.7 (Uniform row-norm alignment)**.** *Under* (1)–(2)*,*

$$\sup_{W \in \mathcal{M}_{\mathrm{row}}^X} D_{\text{avg-bdry}}(W) = o_{\mathbb{P}}(1). \tag{57}$$

In fact, the stronger uniform pairwise statement

$$\sup_{W \in \mathcal{M}_{\mathrm{row}}^X} \max_{i \ne j} \|u_{ij}^W - u_{ij}^\mu\|_2 = o_{\mathbb{P}}(1) \tag{58}$$

holds.

Proof. Work on $\mathcal{E}_d^{\mathrm{det}}$ from Proposition C.4. Apply Proposition C.6 with $h = \varepsilon_d$. Since $k/d \to \alpha \in (0, \infty)$, we have $k \to \infty$, and therefore, uniformly over all row-norm minimizers,

$$\max_i \|w_i - m_i\|_2 = o(1).$$

Combining this with (H1) gives

$$\Delta_d^{\mathrm{row}} := \sup_{W \in \mathcal{M}_{\mathrm{row}}^X} \max_i \|w_i - \mu_i\|_2 = o(1)$$

on $\mathcal{E}_d^{\mathrm{det}}$. Moreover, $\inf_{i \ne j} \|\mu_i - \mu_j\|_2$ is bounded below by a positive constant on the same event. Hence

$$\|(w_i - w_j) - (\mu_i - \mu_j)\|_2 \le 2\Delta_d^{\mathrm{row}}$$

uniformly over $i \ne j$ and $W \in \mathcal{M}_{\mathrm{row}}^X$. Applying Lemma C.5 proves (58), and averaging over ordered pairs gives (57). □

C.4 Max-norm Representatives Have a Positive Angle Gap

The max-norm proof has two ingredients. First, optimality imposes a coordinate cap of order $d^{-1/2}$ on every pairwise row difference. Second, feasibility forces every such difference to have a large projection onto the empirical mean-pair direction. The clipped minimum-energy problem below shows that these two requirements are incompatible with near-perfect angular alignment.

C.4.1 The Clipped Minimum-energy Problem

Proposition C.8 (Exact clipped problem)**.** *Fix $u \in \mathbb{S}^{d-1}$ and $a, b > 0$. Define*

$$\mathfrak{m}_d(u; a, b)^2 := \min\left\{\|x\|_2^2 : x^\top u \ge b,\ \|x\|_\infty \le \frac{a}{\sqrt{d}}\right\}. \tag{59}$$

Let $z_\ell := \sqrt{d}\,|u_\ell|$. If

$$0 < b < a\frac{1}{d}\sum_{\ell=1}^d z_\ell, \tag{60}$$

then the minimizer is unique and has coordinates

$$x_\ell = \frac{a}{\sqrt{d}}\operatorname{sign}(u_\ell)\min\{\tau z_\ell, 1\}, \qquad \ell = 1, \dots, d, \tag{61}$$

where $\operatorname{sign}(0) := 0$ *and $\tau > 0$ is the unique solution of*

$$H_u(\tau) = \frac{b}{a}.$$

Moreover,

$$\mathfrak{m}_d(u; a, b)^2 = a^2 J_u(\tau). \tag{62}$$

Proof. We first verify that the feasible set is nonempty. Define the coordinate-saturated vector $\widehat{x} \in \mathbb{R}^d$ by

$$\widehat{x}_\ell := \frac{a}{\sqrt{d}} \operatorname{sign}(u_\ell).$$

Its projection onto u is

$$\widehat{x}^\top u = \frac{a}{\sqrt{d}} \sum_{\ell=1}^{d} |u_\ell| = \frac{a}{d} \sum_{\ell=1}^{d} z_\ell > b,$$

where the strict inequality follows from (60). Choose c such that $b/(\widehat{x}^\top u) < c < 1$. Then $c\widehat{x}$ satisfies the projection constraint strictly and lies strictly inside the coordinate box. The feasible set is thus nonempty. It is also closed and bounded, so a minimizer exists. Since the objective $x \mapsto \|x\|_2^2$ is strictly convex and the feasible set is convex, this minimizer is unique.

Let $x^\star$ denote the unique minimizer. We next show that every nonzero coordinate of $x^\star$ has the same sign as the corresponding coordinate of u. Given any feasible x, define $x^\sharp$ by

$$x_\ell^\sharp := \begin{cases} \operatorname{sign}(u_\ell)|x_\ell|, & u_\ell \neq 0, \\ 0, & u_\ell = 0. \end{cases}$$

Then $\|x^\sharp\|_\infty \le \|x\|_\infty$, $\|x^\sharp\|_2 \le \|x\|_2$, and

$$(x^\sharp)^\top u = \sum_{\ell=1}^{d} |x_\ell||u_\ell| \ge \sum_{\ell=1}^{d} x_\ell u_\ell = x^\top u.$$

Applying this construction to $x^\star$ shows that $x^\sharp$ is feasible and has objective value no larger than that of $x^\star$. By optimality and uniqueness, $x^\star = x^\sharp$. Hence $x_\ell^\star$ has the sign of u_ℓ when $u_\ell \neq 0$, and $x_\ell^\star = 0$ when $u_\ell = 0$.

The projection constraint must be active at the minimizer. Indeed, if $(x^\star)^\top u > b$, then

$$c := \frac{b}{(x^\star)^\top u}$$

satisfies $0 < c < 1$. The vector $cx^\star$ obeys $\|cx^\star\|_\infty \le a/\sqrt{d}$ and $(cx^\star)^\top u = b$, while

$$\|cx^\star\|_2^2 = c^2 \|x^\star\|_2^2 < \|x^\star\|_2^2,$$

contradicting the optimality of $x^\star$. Therefore,

$$(x^\star)^\top u = b.$$

Because of the sign alignment, write

$$y_\ell := \frac{\sqrt{d}}{a} |x_\ell^\star|, \qquad 0 \le y_\ell \le 1.$$

Using $|u_\ell| = z_\ell/\sqrt{d}$, the active projection constraint becomes

$$b = \sum_{\ell=1}^{d} |x_\ell^\star||u_\ell| = \frac{a}{d} \sum_{\ell=1}^{d} z_\ell y_\ell,$$

or equivalently,

$$\sum_{\ell=1}^{d} z_\ell y_\ell = \frac{db}{a}.$$

Moreover,

$$\|x^\star\|_2^2 = \frac{a^2}{d} \sum_{\ell=1}^{d} y_\ell^2.$$

Therefore, finding $x^\star$ is equivalent to solving the convex optimization problem

$$\min_{0 \le y_\ell \le 1} \frac{1}{2} \sum_{\ell=1}^{d} y_\ell^2 \qquad \text{subject to} \qquad \sum_{\ell=1}^{d} z_\ell y_\ell \ge \frac{db}{a}. \tag{63}$$

The strict inequality in (60) implies that the right-hand side of the constraint is strictly smaller than $\sum_\ell z_\ell$, so the reduced problem has a strictly feasible point. Hence the KKT conditions are necessary and sufficient.

Let $\tau \geq 0$ be the multiplier associated with the constraint in (63). For fixed τ, the Lagrangian minimization over the box $[0,1]^d$ separates across coordinates:

$$y_\ell \in \operatorname*{arg\,min}_{0 \leq y \leq 1} \left\{ \frac{1}{2} y^2 - \tau z_\ell y \right\}. \tag{64}$$

The unconstrained minimizer of the scalar objective is $y = \tau z_\ell$. Projecting this value onto $[0,1]$ gives

$$y_\ell = \min\{\tau z_\ell, 1\}. \tag{65}$$

The multiplier must satisfy $\tau > 0$. If $\tau = 0$, then (64) gives $y_\ell = 0$ for every ℓ, which violates the constraint because $b > 0$. Since $\tau > 0$, complementary slackness implies that the projection constraint is active:

$$\sum_{\ell=1}^{d} z_\ell \min\{\tau z_\ell, 1\} = \frac{db}{a}.$$

After dividing by d, this is exactly

$$H_u(\tau) = \frac{1}{d} \sum_{\ell=1}^{d} z_\ell \min\{\tau z_\ell, 1\} = \frac{b}{a}.$$

It remains to verify that this equation has a unique solution. The function H_u is continuous and nondecreasing on $[0, \infty)$, with

$$H_u(0) = 0, \qquad \lim_{\tau \to \infty} H_u(\tau) = \frac{1}{d} \sum_{\ell=1}^{d} z_\ell.$$

The condition (60) places b/a strictly between these two values, so existence follows from continuity. To prove uniqueness, take $0 \leq \tau_1 < \tau_2$ and suppose that

$$H_u(\tau_1) < \frac{1}{d} \sum_{\ell=1}^{d} z_\ell.$$

Then at least one coordinate with $z_\ell > 0$ is not saturated at τ_1, meaning $\tau_1 z_\ell < 1$. For that coordinate,

$$\min\{\tau_2 z_\ell, 1\} > \min\{\tau_1 z_\ell, 1\},$$

while every other coordinate contributes a nondecreasing term. Thus $H_u(\tau_2) > H_u(\tau_1)$. Hence H_u is strictly increasing before reaching its limiting value, and the equation $H_u(\tau) = b/a$ has a unique solution.

Substituting (65) into the sign-aligned representation of $x^\star$ gives

$$x_\ell^\star = \frac{a}{\sqrt{d}} \operatorname{sign}(u_\ell) \min\{\tau z_\ell, 1\},$$

which proves (61). Finally,

$$\begin{aligned} \mathfrak{m}_d(u; a, b)^2 &= \|x^\star\|_2^2 \\ &= \frac{a^2}{d} \sum_{\ell=1}^{d} \min\{\tau z_\ell, 1\}^2 \\ &= a^2 \frac{1}{d} \sum_{\ell=1}^{d} \min\{\tau^2 z_\ell^2, 1\} \\ &= a^2 J_u(\tau), \end{aligned}$$

which proves (62). □

Lemma C.9 (Uniform asymptotics of the clipped problem)**.** *There exist constants $\delta_0, C_0 > 0$ such that the following holds. Suppose $u \in \mathbb{S}^{d-1}$ satisfies*

$$\sup_{0\le\tau\le1} |H_u(\tau) - A_{\max}(\tau)| + \sup_{0\le\tau\le1} |J_u(\tau) - B_{\max}(\tau)| \le \delta \tag{66}$$

with $0 < \delta \le \delta_0$, and suppose

$$|a - \sqrt{2\pi}| + |b - \sqrt{2}| \le \delta. \tag{67}$$

Then (60) *holds, the solution in Proposition C.8 satisfies $\tau \in (0, 1)$, and*

$$\frac{1}{\mathfrak{m}_d(u; a, b)} \le C_{\max} + C_0\delta, \tag{68}$$

$$\frac{b}{\mathfrak{m}_d(u; a, b)} \le \kappa_{\max} + C_0\delta. \tag{69}$$

Proof. Throughout the proof, $C > 0$ denotes a constant independent of d, u, a, b, and δ, and its value may change from line to line. Recall that $\tau_* > 0$ is defined by

$$A_{\max}(\tau_*) = \frac{1}{\sqrt{\pi}}. \tag{70}$$

We first compare the required normalized projection b/a with this reference value. By taking δ_0 small enough, (67) ensures that $a \ge \sqrt{2\pi}/2$. Hence

$$\begin{aligned}\left|\frac{b}{a} - \frac{1}{\sqrt{\pi}}\right| &= \left|\frac{b}{a} - \frac{\sqrt{2}}{\sqrt{2\pi}}\right| \\ &\le \frac{|b - \sqrt{2}|}{a} + \sqrt{2}\left|\frac{1}{a} - \frac{1}{\sqrt{2\pi}}\right| \\ &= \frac{|b - \sqrt{2}|}{a} + \frac{\sqrt{2}\,|a - \sqrt{2\pi}|}{a\sqrt{2\pi}} \le C\delta.\end{aligned} \tag{71}$$

Thus the right-hand side of the equation $H_u(\tau) = b/a$ is within $O(\delta)$ of $A_{\max}(\tau_*)$.

We next show that the clipped problem is feasible and that its multiplier satisfies $\tau < 1$. Since $\tau_* < 1$ and $A_{\max}$ is strictly increasing, the constant

$$g_* := A_{\max}(1) - A_{\max}(\tau_*) = A_{\max}(1) - \frac{1}{\sqrt{\pi}}$$

is strictly positive. By (66),

$$H_u(1) \ge A_{\max}(1) - \delta = \frac{1}{\sqrt{\pi}} + g_* - \delta.$$

On the other hand, (71) gives

$$\frac{b}{a} \le \frac{1}{\sqrt{\pi}} + C\delta.$$

After decreasing δ_0 if necessary, these two bounds imply

$$0 < \frac{b}{a} < H_u(1).$$

Because

$$H_u(1) = \frac{1}{d}\sum_{\ell=1}^{d} z_\ell \min\{z_\ell, 1\} \le \frac{1}{d}\sum_{\ell=1}^{d} z_\ell,$$

we obtain

$$0 < b < a\frac{1}{d}\sum_{\ell=1}^{d} z_\ell.$$

This is exactly (60). Moreover, $H_u(0) = 0$ and $H_u(1) > b/a$, so Proposition C.8 gives a unique solution $\tau \in (0, 1)$ of

$$H_u(\tau) = \frac{b}{a}. \tag{72}$$

We now show that this solution is close to τ_*. Using (72), (66), (71), and (70), we obtain

$$\begin{aligned} |A_{\max}(\tau) - A_{\max}(\tau_*)| &\le |A_{\max}(\tau) - H_u(\tau)| + |H_u(\tau) - A_{\max}(\tau_*)| \\ &= |A_{\max}(\tau) - H_u(\tau)| + \left| \frac{b}{a} - \frac{1}{\sqrt{\pi}} \right| \le C\delta. \end{aligned} \tag{73}$$

For $s \in (0, 1]$, differentiation under the expectation gives

$$A'_{\max}(s) = \mathbb{E}\left[G^2 \mathbf{1}_{\{s|G|\le 1\}}\right].$$

Since $s \le 1$, the event $\{|G| \le 1\}$ is contained in $\{s|G| \le 1\}$. Therefore

$$A'_{\max}(s) \ge c_A, \qquad c_A := \mathbb{E}\left[G^2 \mathbf{1}_{\{|G|\le 1\}}\right] > 0.$$

The mean value theorem and (73) now give

$$c_A|\tau - \tau_*| \le |A_{\max}(\tau) - A_{\max}(\tau_*)| \le C\delta,$$

and hence

$$|\tau - \tau_*| \le C\delta. \tag{74}$$

We next lower-bound $J_u(\tau)$. The first assumption of the lemma gives

$$J_u(\tau) \ge B_{\max}(\tau) - \delta.$$

The function $B_{\max}$ is Lipschitz on $[0, 1]$. Indeed, for $s, t \in [0, 1]$,

$$\begin{aligned} |B_{\max}(s) - B_{\max}(t)| &\le \mathbb{E}\left[\left|\min\{s^2G^2, 1\} - \min\{t^2G^2, 1\}\right|\right] \\ &\le |s^2 - t^2|\mathbb{E}[G^2] \le 2|s - t|. \end{aligned}$$

Combining this bound with (74) gives

$$B_{\max}(\tau) \ge B_{\max}(\tau_*) - C\delta.$$

Therefore,

$$J_u(\tau) \ge B_{\max}(\tau_*) - C\delta. \tag{75}$$

The assumption on a also gives

$$a^2 \ge \left(\sqrt{2\pi} - \delta\right)^2 \ge 2\pi - C\delta. \tag{76}$$

Using the exact identity from Proposition C.8,

$$\mathfrak{m}_d(u; a, b)^2 = a^2 J_u(\tau),$$

together with (75) and (76), we obtain

$$\begin{aligned} \mathfrak{m}_d(u; a, b)^2 &\ge (2\pi - C\delta)\left(B_{\max}(\tau_*) - C\delta\right) \\ &\ge 2\pi B_{\max}(\tau_*) - C\delta. \end{aligned} \tag{77}$$

Set

$$M_* := 2\pi B_{\max}(\tau_*) > 0.$$

By taking δ_0 smaller if necessary, (77) implies

$$\mathfrak{m}_d(u; a, b)^2 \ge M_* - C\delta \ge \frac{M_*}{2}.$$

Consequently,

$$\begin{aligned} \frac{1}{\mathfrak{m}_d(u; a, b)} &\le \frac{1}{\sqrt{M_* - C\delta}} \\ &\le \frac{1}{\sqrt{M_*}} + C\delta \\ &= C_{\max} + C\delta, \end{aligned}$$

where the last equality follows from

$$C_{\max} = (2\pi B_{\max}(\tau_*))^{-1/2}.$$

After enlarging the constant, this proves (68).

Finally, (67) gives $b \le \sqrt{2} + \delta$. Therefore,

$$\begin{aligned}\frac{b}{\mathfrak{m}_d(u; a, b)} &\le (\sqrt{2} + \delta)(C_{\max} + C_0\delta) \\ &\le \sqrt{2}\, C_{\max} + C\delta \\ &= \kappa_{\max} + C\delta,\end{aligned}$$

because $\kappa_{\max} = \sqrt{2}\, C_{\max}$. Enlarging C_0 once more proves (69). □

C.4.2 Application to Max-norm Minimizers

Proposition C.10 (Max-norm boundary separation)**.** *Under* (1)–(2)*,*

$$\inf_{W \in \mathcal{M}^X_{\max}} D_{\text{avg-bdry}}(W) \ge \delta_{\max} - o_{\mathbb{P}}(1). \tag{78}$$

In fact, on the same high-probability event, every ordered pair of every max-norm minimizer has boundary-normal distance at least $\delta_{\max} - o(1)$*.*

Proof. Work on $\mathcal{E}_d^{\text{det}}$ from Proposition C.4. Fix $W \in \mathcal{M}^X_{\max}$ and set $v_{ij} := w_i - w_j$. By (H5),

$$\|v_{ij}\|_\infty \le 2r^*_{\max}(X) \le \frac{a_d}{\sqrt{d}}, \qquad a_d := \sqrt{2\pi} + \varepsilon_d. \tag{79}$$

Averaging the constraints $(w_i - w_j)^\top x_{i,s} \ge 1$ over class i gives $v_{ij}^\top m_i \ge 1$. Applying the reverse constraints for class j gives $v_{ij}^\top m_j \le -1$. Therefore

$$v_{ij}^\top (m_i - m_j) \ge 2.$$

By (H2), uniformly over $i \ne j$,

$$\|m_i - m_j\|_2^2 = \|m_i\|_2^2 + \|m_j\|_2^2 - 2m_i^\top m_j \le 2 + o(1).$$

Hence

$$v_{ij}^\top u_{ij}^m \ge \frac{2}{\|m_i - m_j\|_2} \ge b_d, \qquad b_d := \sqrt{2} - o(1),$$

where the error is deterministic on $\mathcal{E}_d^{\text{det}}$ and uniform over all pairs.

Fix $i \ne j$, and abbreviate $u = u_{ij}^m$, $v = v_{ij}$, $a = a_d$, and $b = b_d$. Since $v^\top u \ge b > 0$, define

$$y := \frac{b}{v^\top u} v.$$

Then $y^\top u = b$, and because $b/(v^\top u) \le 1$, (79) gives $\|y\|_\infty \le a/\sqrt{d}$. Thus y is feasible for (59), and

$$\mathfrak{m}_d(u; a, b) \le \|y\|_2 = \frac{b\|v\|_2}{v^\top u}.$$

Equivalently,

$$\frac{v^\top u}{\|v\|_2} \le \frac{b}{\mathfrak{m}_d(u; a, b)}.$$

By (H6) and Lemma C.9, the right-hand side is at most $\kappa_{\max} + o(1)$, uniformly over all ordered pairs. Since $\kappa_{\max} < 1$,

$$\left\| \frac{v_{ij}}{\|v_{ij}\|_2} - u_{ij}^m \right\|_2 = \sqrt{2 - 2\frac{v_{ij}^\top u_{ij}^m}{\|v_{ij}\|_2}} \ge \delta_{\max} - o(1).$$

Finally, (H4) and the reverse triangle inequality imply

$$\left\| \frac{v_{ij}}{\|v_{ij}\|_2} - u_{ij}^\mu \right\|_2 \ge \delta_{\max} - o(1),$$

uniformly over all pairs and all max-norm minimizers. Averaging proves (78). □

C.5 Spectral Polar Geometry

Let

$$P := I_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top, \qquad M_m := \begin{bmatrix} m_1^\top \\ \vdots \\ m_k^\top \end{bmatrix}, \qquad C_m := PM_m, \qquad r_d := \min\{k-1, d\}.$$

On the full reduced-rank event, write

$$C_m = U_m \Sigma_m V_m^\top, \qquad U_m \in \mathbb{R}^{k\times r_d}, \quad V_m \in \mathbb{R}^{d\times r_d},$$

and define the reduced polar factor and its positive-semidefinite companion by

$$Q_m := U_m V_m^\top = C_m (C_m^\top C_m)^{\dagger 1/2}, \qquad S_m := Q_m C_m^\top = (C_m C_m^\top)^{1/2}.$$

Write $q_i^\top$ for the i-th row of Q_m.

Lemma C.11 (Quantitative Gaussian hard-edge count at the square aspect ratio)**.** *Let $p = p(d)$ satisfy $p/d \to 1$, let $H_d \in \mathbb{R}^{p\times d}$ have independent $\mathcal{N}(0,1)$ entries, and set $A_d := H_d/\sqrt{d}$ and $n_d := \min\{p, d\}$. There are universal constants $T, c, C > 0$ such that, for all sufficiently large d and every $2/n_d \le u \le T$, an event $\mathcal{H}_d(u)$ satisfies*

$$\mathbb{P}\big(\mathcal{H}_d(u)^c\big) \le C\big(1 + \log(T/u)\big)e^{-cdu^2},$$

and, on this event,

$$\frac{1}{n_d}\#\{\ell : \sigma_\ell(A_d) \le t\} \le 8t, \qquad u \le t \le T.$$

Moreover, A_d has rank n_d almost surely.

Proof. Write $n := n_d$ and take $T = 1/16$. Since $n/d \to 1$, for all sufficiently large d we have $n/d \ge 9/10$. Let B be a fixed $n \times n$ coordinate submatrix of H_d, obtained by retaining a fixed set of n rows when $p \ge d$ and a fixed set of n columns when $p < d$. If $p \ge d$, then $H_d^\top H_d = B^\top B + R^\top R \succeq B^\top B$ for the matrix R of deleted rows, while if $p < d$, then $H_d H_d^\top = BB^\top + RR^\top \succeq BB^\top$ for the matrix R of deleted columns. Hence, by monotonicity of eigenvalues under the positive-semidefinite order,

$$\sigma_j(H_d) \ge \sigma_j(B), \qquad 1 \le j \le n.$$

Fix $t \in [1/n, T]$ and put $\ell := \lceil 3nt \rceil$. For all sufficiently large n, the choice $T = 1/16$ ensures $\ell < n$. Let F be the $n \times (n-\ell)$ matrix obtained by retaining a fixed set of $n - \ell$ columns of B. Apply Vershynin (2012, Corollary 5.35) to F with row dimension $N = n$, column dimension $n - \ell$, and deviation parameter $s := t\sqrt{n}/4$. Then, except with probability at most

$$2e^{-s^2/2} = 2e^{-nt^2/32},$$

we have

$$\sigma_{\min}(F) \ge \sqrt{n} - \sqrt{n-\ell} - s = \sqrt{n} - \sqrt{n-\ell} - \frac{t\sqrt{n}}{4}.$$

Using

$$\sqrt{n} - \sqrt{n-\ell} = \frac{\ell}{\sqrt{n} + \sqrt{n-\ell}} \ge \frac{\ell}{2\sqrt{n}} \ge \frac{3}{2}t\sqrt{n},$$

we obtain

$$\sigma_{\min}(F) \ge \frac{5}{4}t\sqrt{n} > t\sqrt{d},$$

where the last inequality holds for all sufficiently large d because $n/d \to 1$.

Writing $B = [F \;\; R_F]$, we have $BB^\top = FF^\top + R_F R_F^\top \succeq FF^\top$. Therefore the largest $n - \ell$ eigenvalues of $BB^\top$ are at least the $n - \ell$ positive eigenvalues of $FF^\top$, and hence

$$\sigma_j(B) > t\sqrt{d}, \qquad 1 \le j \le n - \ell.$$

Together with $\sigma_j(H_d) \geq \sigma_j(B)$, this gives

$$\frac{1}{n}\#\{j : \sigma_j(A_d) \leq t\} \leq \frac{\ell}{n} \leq 3t + \frac{1}{n} \leq 4t.$$

Now fix $u \in [2/n, T]$ and take the dyadic grid $t_j = 2^{-j}T$ down to the first grid point lying in $[u/2, u]$. Every such grid point is at least $u/2 \geq 1/n$, and the number of grid points is at most $C(1 + \log(T/u))$. Applying the preceding fixed-threshold estimate at every grid point and taking a union bound, the probability that any of these estimates fails is at most

$$C\big(1 + \log(T/u)\big)e^{-cdu^2},$$

because the smallest grid point is at least $u/2$ and $n/d \to 1$. On the complementary event, for every $t \in [u, T]$ choose a dyadic grid point $t_j \in [t, 2t]$; monotonicity of the counting function gives

$$\frac{1}{n}\#\{\ell : \sigma_\ell(A_d) \leq t\} \leq \frac{1}{n}\#\{\ell : \sigma_\ell(A_d) \leq t_j\} \leq 4t_j \leq 8t.$$

Finally, A_d has rank n almost surely because the rank-deficient matrices form a Lebesgue-null algebraic set. □

Proposition C.12 (All-aspect spectral input)**.** *There are deterministic $\eta_d \downarrow 0$, constants $c, C > 0$, and events $\mathcal{E}_d^{\mathrm{sp}}$ such that*

$$\mathbb{P}((\mathcal{E}_d^{\mathrm{sp}})^c) \leq C \exp\Big[-c\sqrt{d \log(2N(d)(d+1))}\Big] \tag{80}$$

and the following statements hold on $\mathcal{E}_d^{\mathrm{sp}}$:

(S1) C_m has rank r_d, $PQ_m = Q_m$, $\|Q_m\|_{\mathrm{op}} = 1$, and

$$\left|\frac{\|C_m\|_*}{k} - s_\alpha\right| \leq \eta_d, \qquad \inf_{i \neq j} \|q_i - q_j\|_2 \geq c_\alpha > 0.$$

(S2) Uniformly over $i \neq j$,

$$\max_i \|m_i - \mu_i\|_2 + \|\bar{m}\|_2 \leq \eta_d, \qquad \bar{m} := \frac{1}{k}\sum_i m_i,$$

$$\max_i |S_m[i,i] - s_\alpha| + \max_{i \neq j} |S_m[i,j]| \leq \eta_d, \qquad \max_{i \neq j, s} |(q_i - q_j)^\top e_{i,s}| \leq \eta_d.$$

(S3) The reduced polar margin

$$\rho_X := \min_{i \neq j, s} (q_i - q_j)^\top x_{i,s}$$

satisfies

$$|\rho_X - s_\alpha| \leq C\eta_d, \qquad \rho_X > 0.$$

(S4) The reduced polar geometry satisfies

$$|D_{\text{avg-bdry}}(Q_m) - \delta_{\mathrm{sp}}(\alpha)| \leq \eta_d. \tag{81}$$

(S5) If $\alpha \neq 1$, all positive singular values of C_m lie in a deterministic compact interval bounded away from zero:

$$c_\alpha \leq \sigma_{r_d}(C_m) \leq \sigma_1(C_m) \leq C_\alpha. \tag{82}$$

If $\alpha = 1$, there are $t_0, C_0 > 0$ such that

$$\frac{1}{r_d}\#\{\ell : \sigma_\ell(C_m) \leq t\} \leq C_0(t + \eta_d), \qquad \eta_d \leq t \leq t_0. \tag{83}$$

Proof. Because $m_i = \mu_i + r\bar{z}_i$ is Gaussian with covariance $(1+r^2/n_i)I_d/d$, the rows are independent and admit the exact representation

$$M_m = D_d G/\sqrt{d}, \qquad D_d := \operatorname{diag}\Big(\sqrt{1+r^2/n_1}, \dots, \sqrt{1+r^2/n_k}\Big),$$

where G has independent standard Gaussian entries. Since $\|D_d - I_k\|_{\mathrm{op}} \le Cr_0^2/n_{\min} = o(1)$ and $\|G/\sqrt{d}\|_{\mathrm{op}} = O(1)$ with exponentially high probability, the comparator $C_0 := PG/\sqrt{d}$ obeys

$$\|C_m - C_0\|_{\mathrm{op}} \le \zeta_d, \qquad \zeta_d \downarrow 0. \tag{84}$$

After a fixed orthogonal change of basis on $\mathbb{R}^k$, the positive singular values of C_0 have the same law as those of a $(k-1)\times d$ standard Gaussian matrix divided by $\sqrt{d}$. Marchenko–Pastur convergence therefore gives $\|C_0\|_*/k \to s_\alpha$. The normalized nuclear norm is $O(d^{-1})$-Lipschitz in the Gaussian entries, so Gaussian concentration and the deterministic expectation bias place this convergence under a deterministic envelope with failure probability bounded as in (80). The rank assertion is almost sure, and Weyl's inequality yields (82) when $\alpha \neq 1$.

At $\alpha = 1$, put $L_d := \log(2N(d)(d+1))$ and $u_d := (L_d/d)^{1/4}$. Then $u_d \to 0$, $u_d \ge 2/r_d$ eventually, and Lemma C.11 gives failure at most $C\exp(-c\sqrt{dL_d})$ after absorbing its logarithmic prefactor. Weyl's inequality and (84) transfer the count from C_0 to C_m by replacing t with $t + \zeta_d$. Enlarge the common envelope so that $\eta_d \ge \max\{u_d, \zeta_d\}$ and take $t_0 < T/2$; then (83) follows with precisely the probability scale in (80).

Let $S_0 := (C_0 C_0^\top)^{1/2}$. Conditionally on its singular values, the left singular frame of C_0 is Haar in $\mathbf{1}^\perp$ and independent of those singular values. Haar concentration for the quadratic forms defining $S_0[i,j]$, followed by a union bound over k^2 entries, gives

$$\max_i \left| S_0[i,i] - \frac{\|C_0\|_*}{k} \right| + \max_{i\neq j} |S_0[i,j]| \le \chi_d, \qquad \chi_d \downarrow 0.$$

The gap-free square-root inequality for positive semidefinite matrices gives

$$\|S_m - S_0\|_{\mathrm{op}} \le \|C_m C_m^\top - C_0 C_0^\top\|_{\mathrm{op}}^{1/2} \le C\zeta_d^{1/2}.$$

This proves the entrywise bounds in (S2) without dividing by the smallest singular value. The empirical-mean bound is part of Proposition C.4, and the Gaussian average $k^{-1}\sum_i m_i$ gives $\|\bar{m}\|_2 = o(1)$ under the same probability envelope.

Conditionally on the empirical means, Q_m is deterministic and the residuals are centered Gaussian. Since $\|q_i - q_j\|_2 \le \sqrt{2}$, Gaussian tails and a union bound over at most N^2 indices give the residual bound in (S2). Using that row i of C_m is $(m_i - \bar{m})^\top$, we obtain

$$(q_i - q_j)^\top m_i = S_m[i,i] - S_m[j,i] + (q_i - q_j)^\top \bar{m} = s_\alpha + O(\eta_d).$$

Adding the residual bound proves (S3). The same identity and $\max_{i,s}\|x_{i,s}\|_2 = O(1)$ give the uniform lower bound on $\|q_i - q_j\|_2$.

It remains to identify the angular constant. The centering identities $\mathbf{1}^\top Q_m = \mathbf{0}$ and $\|Q_m\|_{\mathrm{F}}^2 = r_d$ give

$$\frac{1}{k(k-1)}\sum_{i\neq j}\|q_i - q_j\|_2^2 = \frac{2r_d}{k-1} \longrightarrow \frac{2\rho_\alpha}{\alpha}. \tag{85}$$

If $\alpha < 1$, then $r_d = k-1$ and $Q_m Q_m^\top = P$, so every pair norm is exactly $\sqrt{2}$. For $\alpha > 1$, let Q_0 be the reduced polar factor of C_0 and put $\Pi_m := Q_m Q_m^\top$ and $\Pi_0 := Q_0 Q_0^\top$. The left singular frame of C_0, but not in general that of $C_m = PD_dG/\sqrt{d}$, is Haar in $\mathbf{1}^\perp$ when the class counts differ. The positive singular-value lower bounds and (84) give

$$\|(I - \Pi_0)\Pi_m\|_{\mathrm{op}} \le \|C_m - C_0\|_{\mathrm{op}}\|C_m^\dagger\|_{\mathrm{op}} \le C_\alpha\zeta_d,$$

and the analogous estimate with 0 and m exchanged. Consequently, $\|\Pi_m - \Pi_0\|_{\mathrm{op}} \le C_\alpha\zeta_d$. Haar-projector concentration applies to Π_0 and yields $(e_i - e_j)^\top\Pi_0(e_i - e_j) = 2/\alpha + o(1)$ uniformly over pairs on the same probability scale. Since each such quadratic form changes by at most $2\|\Pi_m - \Pi_0\|_{\mathrm{op}}$,

we obtain $\|q_i - q_j\|_2 = \sqrt{2/\alpha} + o(1)$ uniformly. At $\alpha = 1$, no perturbation of the polar factor at the hard edge is needed for this step. The deterministic bounds $\|q_i - q_j\|_2 \le \sqrt{2}$ and (85) give

$$\frac{1}{k(k-1)} \sum_{i \neq j} \left| \|q_i - q_j\|_2 - \sqrt{2} \right| \le \frac{1}{\sqrt{2}} \left(2 - \frac{2r_d}{k-1} \right) \longrightarrow 0.$$

Thus the required pair-norm convergence holds in averaged L^1 at $\alpha = 1$.

Furthermore,

$$(q_i - q_j)^\top (\mu_i - \mu_j) = S_m[i,i] + S_m[j,j] - S_m[i,j] - S_m[j,i] + o(1) = 2s_\alpha + o(1),$$

while $\|\mu_i - \mu_j\|_2 = \sqrt{2} + o(1)$ uniformly. Consequently, the pairwise cosine converges in averaged L^1 to

$$\frac{2s_\alpha}{\sqrt{2\rho_\alpha/\alpha}\sqrt{2}} = \frac{m_{1/2}(\alpha)}{\sqrt{\rho_\alpha}} = \kappa_{\mathrm{sp}}(\alpha).$$

Uniform continuity of $x \mapsto \sqrt{2 - 2x}$ on $[-1, 1]$ proves (S4). Collecting the deterministic envelopes proves the proposition. □

Lemma C.13 (Weighted trace gap and hard-edge truncation)**.** *Let $C = U\Sigma V^\top$ have rank r, let $Q = UV^\top$, and let B satisfy $\|B\|_{\mathrm{op}} \le 1$ and $\mathrm{rank}(B) \le r$. Put $g_\ell := u_\ell^\top B v_\ell$ and $a_\ell := 1 - g_\ell \in [0, 2]$. Then*

$$\|B - Q\|_{\mathrm{F}}^2 \le 2 \sum_{\ell=1}^{r} a_\ell,$$
$$\frac{1}{r} \sum_{\ell=1}^{r} a_\ell \le 2 \frac{\#\{\ell : \sigma_\ell \le \tau\}}{r} + \frac{1}{\tau r} \sum_{\ell=1}^{r} \sigma_\ell a_\ell, \qquad \tau > 0. \tag{86}$$

Proof. The rank and operator-norm bounds give $\|B\|_{\mathrm{F}}^2 \le r$, so expanding $\|B - Q\|_{\mathrm{F}}^2$ proves the first inequality. For the second inequality, use $a_\ell \le 2$ when $\sigma_\ell \le \tau$ and $a_\ell \le \sigma_\ell a_\ell / \tau$ otherwise. □

Lemma C.14 (Average normalization transfer)**.** *Let $B, Q \in \mathbb{R}^{k \times d}$ have rows b_i, q_i, suppose $b_i - b_j \neq \mathbf{0}$ for $i \neq j$, and assume $\inf_{i \neq j} \|q_i - q_j\|_2 \ge c_0 > 0$. If*

$$E_2 := \frac{1}{k(k-1)} \sum_{i \neq j} \|(b_i - b_j) - (q_i - q_j)\|_2^2,$$

then

$$\frac{1}{k(k-1)} \sum_{i \neq j} \left\| \frac{b_i - b_j}{\|b_i - b_j\|_2} - \frac{q_i - q_j}{\|q_i - q_j\|_2} \right\|_2 \le \frac{4}{c_0} E_2^{1/2} + \frac{8}{c_0^2} E_2.$$

Proof. For a pairwise difference error $e_{ij} \le c_0/2$, the normalization map gives an error at most $4e_{ij}/c_0$. For the remaining pairs, use the trivial bound two and $\mathbf{1}_{\{e_{ij} > c_0/2\}} \le 4e_{ij}^2/c_0^2$, then average and apply Cauchy–Schwarz. □

Proposition C.15 (Exact and near spectral minimizers)**.** *On the events of Proposition C.12, let $W \in \mathcal{F}_X$ satisfy*

$$\|W\|_{\mathrm{op}} \le (1 + \xi) r_{\mathrm{sp}}^*(X), \qquad 0 \le \xi \le 1.$$

Then

$$D_{\text{avg-bdry}}(W) \ge \delta_{\mathrm{sp}}(\alpha) - \begin{cases} C_\alpha (\eta_d + \xi)^{1/2}, & \alpha \neq 1, \\ C (\eta_d + \xi)^{1/4}, & \alpha = 1. \end{cases} \tag{87}$$

In particular, every exact minimizer satisfies

$$\frac{1}{r_d} \left\| \frac{PW}{\|PW\|_{\mathrm{op}}} - Q_m \right\|_{\mathrm{F}}^2 \le \begin{cases} C_\alpha \eta_d, & \alpha \neq 1, \\ C \sqrt{\eta_d}, & \alpha = 1, \end{cases} \tag{88}$$

uniformly over $W \in \mathcal{M}_{\mathrm{sp}}^X$.

Proof. The polar witness Q_m/ρ_X is feasible, so

$$r^*_{\mathrm{sp}}(X) \le \rho_X^{-1}. \tag{89}$$

For every feasible W, summing the class-averaged constraints over ordered pairs gives

$$\langle PW, C_m \rangle \ge k - 1. \tag{90}$$

Set $A := PW$, $a := \|A\|_{\mathrm{op}} > 0$, and $B := A/a$. Centering is an operator-norm contraction and preserves all row differences, while $\mathrm{rank}(B) \le r_d$. Equations (89), (90), and Proposition C.12 give

$$\frac{1}{r_d} \sum_{\ell=1}^{r_d} \sigma_\ell(C_m) \left(1 - u_\ell^\top B v_\ell\right) \le C(\eta_d + \xi).$$

If $\alpha \neq 1$, the positive edge in (82) and Lemma C.13 yield $r_d^{-1}\|B - Q_m\|_{\mathrm{F}}^2 \le C_\alpha(\eta_d + \xi)$. If $\alpha = 1$, set $q := \eta_d + \xi$ and use $\tau = \sqrt{q}$ in (86) together with (83); the result is

$$\frac{1}{r_d}\|B - Q_m\|_{\mathrm{F}}^2 \le C q^{1/2}.$$

The case of non-small q follows by increasing C and using the trivial bound. Taking $\xi = 0$ proves (88).

Because both B and Q_m are centered,

$$\frac{1}{k(k-1)} \sum_{i \neq j} \|(b_i - b_j) - (q_i - q_j)\|_2^2 = \frac{2}{k-1}\|B - Q_m\|_{\mathrm{F}}^2.$$

Feasibility ensures $b_i - b_j \neq \mathbf{0}$, and Proposition C.12 supplies a uniform lower bound on the polar pair norms. Lemma C.14, followed by the reverse triangle inequality and (81), proves (87). □

Proposition C.16 (Spectral-norm boundary separation)**.** *Under Assumptions 2.1 and 2.2,*

$$\inf_{W \in \mathcal{M}_{\mathrm{sp}}^X} D_{\text{avg-bdry}}(W) \ge \delta_{\mathrm{sp}}(\alpha) - o_{\mathbb{P}}(1).$$

The bound holds on the events in Proposition C.12 with the deterministic error in (87) *at* $\xi = 0$*.*

Proof. Apply Proposition C.15 with $\xi = 0$ and take the infimum over the exact minimizer set. □

Proof of the deterministic geometry theorem

Proof of Theorem C.2. Let $\mathcal{G}_d^{\mathrm{det}}$ be the intersection of $\mathcal{E}_d^{\mathrm{det}}$ from Proposition C.4 and $\mathcal{E}_d^{\mathrm{sp}}$ from Proposition C.12. Both complements satisfy a bound of the form

$$C \exp\left[-c\sqrt{d \log(2N(d)(d+1))}\right]. \tag{91}$$

Since $\log(2N(d)(d+1)) \to \infty$, the quantity in (91) is at most $e^{-\sqrt{d}}$ for every sufficiently large d, after increasing the deterministic threshold. The sign construction in Proposition C.4 gives $\mathcal{F}_X \neq \varnothing$ on this event, and compactness gives nonempty minimum-norm sets.

The proof of Proposition C.7 is deterministic on $\mathcal{E}_d^{\mathrm{det}}$ and gives the maximum-pair statement (42) under a deterministic vanishing envelope. Lemma C.3 then gives (43). Proposition C.10 and Lemma C.3 give (45). Proposition C.16 gives (44) for every fixed $\alpha > 0$, using the reduced polar factor when $\alpha < 1$ and the hard-edge truncation when $\alpha = 1$. Choose one deterministic $\omega_d \downarrow 0$ dominating all three errors. Because $\delta_{\mathrm{sp}}(\alpha) > 0$ and $\delta_{\max} > 0$, (46) follows for every sufficiently large d. □

C.6 From Direction-set Geometry to Population Accuracy

We now convert the geometric separation above into ordinary multiclass population accuracy. The lemmas in this subsection are stated for general classifier families so that the same transfer can be reused in the stochastic and AdamProxy analyses.

For every classifier with nonzero pairwise row differences, define

$$\mathcal{E}_{\pi,d}(W) := 1 - \mathrm{Acc}_{\pi,d}(W), \qquad p_{i\to j}(W) := \mathbb{P}\big(w_j^\top X^{\mathrm{te}} \ge w_i^\top X^{\mathrm{te}} \mid Y = i\big).$$

Under the normalized convention in (38), set $\sigma_d := r_{\mathrm{te}}/\sqrt{d}$, and for $i \neq j$ write

$$d_{ij}(W) := \|u_{ij}^W - u_{ij}^\mu\|_2, \qquad s_{ij} := \|\mu_i - \mu_j\|_2, \qquad m_{ij} := \frac{\mu_i + \mu_j}{2}.$$

Lemma C.17 (Pairwise Gaussian error bounds)**.** *For every $i < j$, let $e_{ij}(W) := \{p_{i\to j}(W) + p_{j\to i}(W)\}/2$. Then*

$$\max_{i<j}\{\pi_{i,d}p_{i\to j}(W) + \pi_{j,d}p_{j\to i}(W)\} \le \mathcal{E}_{\pi,d}(W) \le \sum_{i<j}\{\pi_{i,d}p_{i\to j}(W) + \pi_{j,d}p_{j\to i}(W)\}, \quad (92)$$

and hence $\mathcal{E}_{\pi,d}(W) \ge 2\pi_{\min,d}\max_{i<j} e_{ij}(W)$. Moreover, with $c_{ij}(W) := 1 - d_{ij}(W)^2/2$,

$$e_{ij}(W) = \frac{1}{2}\left[\overline{\Phi}\left(\frac{s_{ij}c_{ij}(W)/2 + \langle u_{ij}^W, m_{ij}\rangle}{\sigma_d}\right) + \overline{\Phi}\left(\frac{s_{ij}c_{ij}(W)/2 - \langle u_{ij}^W, m_{ij}\rangle}{\sigma_d}\right)\right] \ge \overline{\Phi}\left(\frac{s_{ij}(c_{ij}(W))_+}{2\sigma_d}\right), \quad (93)$$

where $\overline{\Phi}(a) = \mathbb{P}(Z \ge a)$ for $Z \sim \mathcal{N}(0,1)$.

Proof. Conditionally on $Y = i$, the multiclass error event is the union of the events $\{w_j^\top X^{\mathrm{te}} \ge w_i^\top X^{\mathrm{te}}\}$, up to Gaussian-null ties. The union bound gives the upper half of (92), while any fixed unordered pair gives the lower half. For $u = u_{ij}^W$, $v = u_{ij}^\mu$, $m = m_{ij}$, and $s = s_{ij}$, the identities $\mu_i = m + (s/2)v$ and $\mu_j = m - (s/2)v$ give

$$p_{i\to j}(W) = \overline{\Phi}\left(\frac{u^\top \mu_i}{\sigma_d}\right), \qquad p_{j\to i}(W) = \overline{\Phi}\left(\frac{-u^\top \mu_j}{\sigma_d}\right),$$

which proves the equality in (93). For $q \ge 0$, the even function $z \mapsto [\overline{\Phi}(q+z) + \overline{\Phi}(q-z)]/2$ is minimized at zero, while for $q < 0$ the same average is at least $1/2$. Since $u^\top v = 1 - d_{ij}(W)^2/2$, the final inequality follows. □

Define

$$\tau_d := \max_i \left|\|\mu_i\|_2^2 - 1\right|, \qquad \rho_d := \max_{i\neq j} |\mu_i^\top \mu_j|. \quad (94)$$

Lemma C.18 (Quantitative Gaussian radial event)**.** *There are deterministic $\eta_d \downarrow 0$ and events $\mathcal{R}_d$, depending only on the class means, such that, for every sufficiently large d,*

$$\mathbb{P}(\mathcal{R}_d^c) \le \frac{1}{4}e^{-\sqrt{d}}, \qquad \tau_d + \rho_d \le \eta_d \quad \text{on } \mathcal{R}_d. \quad (95)$$

Proof. Under the normalized convention, $d\|\mu_i\|_2^2$ is chi-square with d degrees of freedom, while conditionally on μ_i, the inner product $\mu_i^\top \mu_j$ is Gaussian with variance $\|\mu_i\|_2^2/d$. Taking a sufficiently large constant C_α and $\eta_d = C_\alpha d^{-1/4}$, standard chi-square and Gaussian tails followed by a union bound over $O(d^2)$ indices give (95). □

For nonempty classifier families $\mathcal{R}_d^{\mathrm{clf}}$ and $\mathcal{C}_d^{\mathrm{clf}}$, put

$$h_d := \sup_{W\in\mathcal{R}_d^{\mathrm{clf}}} \max_{i\neq j} d_{ij}(W), \qquad \Delta_d := \inf_{W\in\mathcal{C}_d^{\mathrm{clf}}} D_{\text{avg-bdry}}(W),$$

and, whenever $\tau_d + \rho_d < 1$, define

$$A_d := \frac{1 - \tau_d - \rho_d}{\sqrt{2(1+\tau_d+\rho_d)}} - h_d\sqrt{1+\tau_d}, \quad (96)$$

$$B_d := \sqrt{\frac{1+\tau_d+\rho_d}{2}}\left(1 - \frac{\Delta_d^2}{2}\right)_+. \quad (97)$$

Proposition C.19 (Finite-dimensional geometry-to-error transfer)**.** *If $A_d > 0$, then*

$$\sup_{W \in \mathcal{R}_d^{\mathrm{clf}}} \mathcal{E}_{\pi,d}(W) \le (k-1)\overline{\Phi}\left(\frac{A_d}{\sigma_d}\right), \qquad \inf_{W \in \mathcal{C}_d^{\mathrm{clf}}} \mathcal{E}_{\pi,d}(W) \ge 2\pi_{\min,d}\overline{\Phi}\left(\frac{B_d}{\sigma_d}\right).$$

Proof. For $v = u_{ij}^{\mu}$, radial control gives

$$v^\top \mu_i \ge \frac{1 - \tau_d - \rho_d}{\sqrt{2(1+\tau_d+\rho_d)}}, \qquad -v^\top \mu_j \ge \frac{1 - \tau_d - \rho_d}{\sqrt{2(1+\tau_d+\rho_d)}}.$$

If $W \in \mathcal{R}_d^{\mathrm{clf}}$, then $\|u_{ij}^W - v\|_2 \le h_d$, so both directed population margins are at least A_d. The upper bound follows from the union bound in Lemma C.17. If $W \in \mathcal{C}_d^{\mathrm{clf}}$, some pair has $d_{ij}(W) \ge \Delta_d$, while $s_{ij}/2 \le \sqrt{(1+\tau_d+\rho_d)/2}$. The midpoint-free lower bound in Lemma C.17 then gives the second inequality. □

Define

$$\mathsf{a}(h) := \frac{1}{\sqrt{2}} - h, \qquad \mathsf{b}(\delta) := \frac{1}{\sqrt{2}}\left(1 - \frac{\delta^2}{2}\right)_+. \tag{98}$$

Theorem C.20 (Geometry-to-accuracy transfer)**.** *Suppose $\tau_d + \rho_d \to 0$, $\limsup_d h_d \le \overline{h}$, and $\liminf_d \Delta_d \ge \underline{\delta} > 0$. If $\mathsf{a}(\overline{h}) > \mathsf{b}(\underline{\delta})$, then under Assumption 2.2,*

$$\frac{\sup_{W \in \mathcal{R}_d^{\mathrm{clf}}} \mathcal{E}_{\pi,d}(W)}{\inf_{W \in \mathcal{C}_d^{\mathrm{clf}}} \mathcal{E}_{\pi,d}(W)} \longrightarrow 0,$$

and therefore, for every sufficiently large d,

$$\inf_{W \in \mathcal{R}_d^{\mathrm{clf}}} \mathrm{Acc}_{\pi,d}(W) > \sup_{W \in \mathcal{C}_d^{\mathrm{clf}}} \mathrm{Acc}_{\pi,d}(W).$$

Proof. Choose constants $0 < b < a$ such that $\mathsf{b}(\underline{\delta}) < b < a < \mathsf{a}(\overline{h})$. Proposition C.19 gives, for all sufficiently large d,

$$\frac{\sup_{\mathcal{R}_d^{\mathrm{clf}}} \mathcal{E}_{\pi,d}}{\inf_{\mathcal{C}_d^{\mathrm{clf}}} \mathcal{E}_{\pi,d}} \le \frac{k-1}{2\pi_{\min,d}} \frac{\overline{\Phi}(a/\sigma_d)}{\overline{\Phi}(b/\sigma_d)}.$$

The standard Mills bounds imply that the right-hand side is at most $Ck\pi_{\min,d}^{-1} e^{-cd}$ for constants $C, c > 0$, because $r_{\mathrm{te}} \le r_{\mathrm{te},0}$. Assumption 2.2 makes this quantity vanish. □

Lemma C.21 (Population-error continuity along a direction set)**.** *Fix d, a linearly separable training set X containing at least one sample from every class in $[k]$, a matrix norm ψ, and $0 \le \varepsilon < \gamma_\psi(X)$. If $d_{\cos}(W_t, \mathcal{U}_{\psi,\varepsilon}^X) \to 0$, then*

$$\limsup_{t\to\infty} \mathcal{E}_{\pi,d}(W_t) \le \sup_{U \in \mathcal{U}_{\psi,\varepsilon}^X} \mathcal{E}_{\pi,d}(U), \qquad \liminf_{t\to\infty} \mathcal{E}_{\pi,d}(W_t) \ge \inf_{U \in \mathcal{U}_{\psi,\varepsilon}^X} \mathcal{E}_{\pi,d}(U).$$

Proof. The norm-normalized section of $\mathcal{U}_{\psi,\varepsilon}^X$ is compact, and its strictly positive empirical margin keeps every pairwise row difference uniformly away from zero. Population error is invariant under positive scaling and continuous on this section because the Gaussian test law gives zero mass to score-tie hyperplanes. Uniform continuity on a compact neighborhood proves the claim. □

Lemma C.22 (Boundary metric under direction-set convergence)**.** *Under the fixed-data assumptions of Lemma C.21, if $d_{\cos}(W_t, \mathcal{U}_{\psi,\varepsilon}^X) \to 0$ with $0 \le \varepsilon < \gamma_\psi(X)$, then*

$$\limsup_{t\to\infty} D_{\text{avg-bdry}}(W_t) \le \sup_{U \in \mathcal{U}_{\psi,\varepsilon}^X} D_{\text{avg-bdry}}(U), \qquad \liminf_{t\to\infty} D_{\text{avg-bdry}}(W_t) \ge \inf_{U \in \mathcal{U}_{\psi,\varepsilon}^X} D_{\text{avg-bdry}}(U).$$

Proof. The same compactness argument as in Lemma C.21 applies because $D_{\text{avg-bdry}}$ is scale invariant and continuous whenever all pairwise row differences are nonzero. □

C.7 Proof of the Deterministic Optimizer Accuracy Ordering

Proof of Theorem 3.1. Work on the event $\mathcal{G}_d^{\mathrm{det}}$ from Theorem C.2. Proposition C.4, items (H1)–(H2) imply $\tau_d + \rho_d = o(1)$ on this event. Take $\mathcal{R}_d^{\mathrm{clf}} = \mathcal{U}_{\mathrm{row}}^X$. Equation (42) gives $h_d = o(1)$. For $\psi \in \{\mathrm{sp}, \max\}$, take $\mathcal{C}_d^{\mathrm{clf}} = \mathcal{U}_\psi^X$. Equations (44) and (45) give $\liminf_d \Delta_d \geq \delta_\psi > 0$, where $\delta_{\mathrm{sp}} = \delta_{\mathrm{sp}}(\alpha)$ and $\delta_{\max}$ is defined in (40). Since

$$\mathsf{a}(0) = \frac{1}{\sqrt{2}} > \frac{1}{\sqrt{2}}\left(1 - \frac{\delta_\psi^2}{2}\right)_+ = \mathsf{b}(\delta_\psi),$$

Theorem C.20 gives, for every sufficiently large d on $\mathcal{G}_d^{\mathrm{det}}$,

$$\sup_{U \in \mathcal{U}_{\mathrm{row}}^X} \mathcal{E}_{\pi,d}(U) < \min_{\psi \in \{\mathrm{sp},\max\}} \inf_{V \in \mathcal{U}_\psi^X} \mathcal{E}_{\pi,d}(V). \tag{99}$$

The full-batch implicit-bias results in Appendix B, together with Lemma B.2, give direction-set convergence to $\mathcal{U}_{\mathrm{row}}^X$ for RowGD and RowGDM, to $\mathcal{U}_{\mathrm{sp}}^X$ for Spectral-GD and exact-SVD Muon, and to $\mathcal{U}_{\max}^X$ for SignGD, Signum, and full-batch Adam under the stated Adam assumptions. Lemma C.21 transfers (99) to all seven trajectories. Since $\mathbb{P}(\mathcal{G}_d^{\mathrm{det}}) \geq 1 - e^{-\sqrt{d}}$, replacing error by one minus accuracy proves (5) with the claimed probability. □

C.8 Proof of the Limiting Classifier Geometry

We prove Proposition 3.2 under the normalized convention (38), under which $\mu_i, z_{i,s} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d/d)$. Since all statements below are invariant under the common rescaling used in (38), they transfer directly to the original normalization.

Let $M_\mu \in \mathbb{R}^{k\times d}$ have rows $\mu_i^\top$, let $C_\mu := PM_\mu$, and write a compact SVD $C_\mu = U_\mu \Sigma_\mu V_\mu^\top$. Define $Q_\mu := U_\mu V_\mu^\top$, and write $q_i^\top$ for its i-th row. Recall from (41) that

$$\kappa_{\mathrm{sp}}(\alpha) = \frac{m_{1/2}(\alpha)}{\sqrt{\alpha \wedge 1}} \in (0,1).$$

In particular, when $\alpha > 1$, $\kappa_{\mathrm{sp}}(\alpha) = m_{1/2}(\alpha)$.

Lemma C.23 (Empirical-to-population polar transfer)**.** *Under Assumptions 2.1 and 2.2,*

$$\frac{1}{r_d}\|Q_m - Q_\mu\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} 0. \tag{100}$$

Proof. By (47), the i-th row of $M_m - M_\mu$ is $r\bar{z}_i^\top$, and the rows are independent centered Gaussian vectors with covariance $r^2 I_d/(n_i d)$. Hence $M_m - M_\mu = D_d^{\mathrm{e}} H/\sqrt{d}$ for a standard Gaussian matrix H and a diagonal matrix satisfying $\|D_d^{\mathrm{e}}\|_{\mathrm{op}} \leq r_0/\sqrt{n_{\min}}$. Since $k/d \to \alpha$ and $n_{\min} \to \infty$,

$$\delta_d := \|C_m - C_\mu\|_{\mathrm{op}} \leq \|M_m - M_\mu\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

On the full reduced-rank event, $\|Q_\mu\|_{\mathrm{op}} = 1$, $\mathrm{rank}(Q_\mu) = r_d$, and $\langle Q_\mu, C_\mu\rangle = \|C_\mu\|_*$. Therefore

$$\begin{aligned} 0 &\leq \|C_m\|_* - \langle Q_\mu, C_m\rangle \\ &\leq \big|\|C_m\|_* - \|C_\mu\|_*\big| + \big|\langle Q_\mu, C_m - C_\mu\rangle\big| \leq 2r_d\delta_d. \end{aligned} \tag{101}$$

Write $C_m = U_m \Sigma_m V_m^\top$, and set $a_\ell := 1 - u_{m,\ell}^\top Q_\mu v_{m,\ell}$. Then (101) gives

$$\frac{1}{r_d}\sum_{\ell=1}^{r_d} \sigma_\ell(C_m) a_\ell \leq 2\delta_d. \tag{102}$$

If $\alpha \neq 1$, Proposition C.12, item (S5), bounds every positive singular value of C_m away from zero, so Lemma C.13 and (102) yield $r_d^{-1}\|Q_m - Q_\mu\|_{\mathrm{F}}^2 = o_{\mathbb{P}}(1)$.

Suppose $\alpha = 1$. Choose a deterministic sequence $\bar{\delta}_d \downarrow 0$ such that $\delta_d \le \bar{\delta}_d$ with probability tending to one, and set $\tau_d := \max\{\eta_d, \sqrt{\bar{\delta}_d}\}$, where η_d is the envelope in Proposition C.12. For all sufficiently large d, $\tau_d \le t_0$, and item (S5) together with Lemma C.13 gives

$$\frac{1}{r_d}\|Q_m - Q_\mu\|_{\mathrm{F}}^2 \le 4C_0(\tau_d + \eta_d) + \frac{4\delta_d}{\tau_d} = o_{\mathbb{P}}(1).$$

This proves (100). □

Proof of Proposition 3.2. We first establish the claimed geometry uniformly over the corresponding exact maximum-margin faces.

For the Row geometry, Proposition C.6 and Proposition C.4, item (H1), give

$$\sup_{W \in \mathcal{M}_{\mathrm{row}}^X} \max_i \|w_i - \mu_i\|_2 = o_{\mathbb{P}}(1).$$

Since $\min_i \|\mu_i\|_2 = 1 - o_{\mathbb{P}}(1)$, Lemma C.5 yields

$$\sup_{W \in \mathcal{M}_{\mathrm{row}}^X} \max_i \|\widehat{w}_i - \widehat{\mu}_i\|_2 = o_{\mathbb{P}}(1). \tag{103}$$

For the Max geometry, put $g := \sqrt{2/\pi}$, let $T_i := \mathrm{sign}(\mu_i)$, and write $t_{\max} := r_{\max}^*(X)$. For $W \in \mathcal{M}_{\max}^X$, set $A := W/t_{\max}$. The sharp radius estimate (114) and the global sign rigidity (119) imply

$$t_{\max} = \frac{1 + o_{\mathbb{P}}(1)}{g\sqrt{d}}, \qquad \sup_{W \in \mathcal{M}_{\max}^X} \frac{\|A - T\|_{\mathrm{F}}^2}{kd} = o_{\mathbb{P}}(1), \tag{104}$$

where T has rows $T_i^\top$. Let $\bar{A} := k^{-1}\sum_i A_i$ and $\bar{T} := k^{-1}\sum_i T_i$. Jensen's inequality, (104), and $\mathbb{E}\|\bar{T}\|_2^2 = d/k = O(1)$ give $\sup_W \|\bar{A}\|_2/\sqrt{d} = o_{\mathbb{P}}(1)$. The class-averaged feasibility constraint (56) then gives, uniformly over i and $W \in \mathcal{M}_{\max}^X$,

$$A_i^\top \mu_i \ge g\sqrt{d} - o_{\mathbb{P}}(\sqrt{d}).$$

Together with $\max_i |\|\mu_i\|_1/\sqrt{d} - g| = o_{\mathbb{P}}(1)$ from (117), this implies $D_i := \|\mu_i\|_1 - A_i^\top \mu_i = o_{\mathbb{P}}(\sqrt{d})$ uniformly. Substituting this bound into (120) and using the uniform Gaussian small-ball estimate (118) yields

$$\sup_{W \in \mathcal{M}_{\max}^X} \max_i \frac{1}{d}\|A_i - T_i\|_2^2 = o_{\mathbb{P}}(1).$$

Since $\|T_i\|_2 = \sqrt{d}$, normalization stability gives

$$\sup_{W \in \mathcal{M}_{\max}^X} \max_i \left\| \widehat{w}_i - \frac{\mathrm{sign}(\mu_i)}{\sqrt{d}} \right\|_2 = o_{\mathbb{P}}(1). \tag{105}$$

For the Spectral geometry, let $B := PW/\|PW\|_{\mathrm{op}}$. Combining (88) with Lemma C.23 gives

$$\sup_{W \in \mathcal{M}_{\mathrm{sp}}^X} \frac{1}{r_d}\|B - Q_\mu\|_{\mathrm{F}}^2 = o_{\mathbb{P}}(1). \tag{106}$$

This proves the population-mean version of the spectral geometry claimed in the proposition.

It remains to identify the mean angles when $\alpha > 1$. For Row, (103) immediately gives a limiting angle of zero. For Max, (105), $\max_i |\|\mu_i\|_2 - 1| = o_{\mathbb{P}}(1)$, and (117) give

$$\sup_{W \in \mathcal{M}_{\max}^X} \max_i \left| \frac{w_i^\top \mu_i}{\|w_i\|_2 \|\mu_i\|_2} - \sqrt{\frac{2}{\pi}} \right| = o_{\mathbb{P}}(1).$$

For Spectral, $\alpha > 1$ implies $r_d = d$. Since C_μ is a centered Gaussian matrix, the left singular frame U_μ is Haar in $\mathbf{1}^\perp$ and independent of its singular values. Haar-projector concentration and the Marchenko–Pastur limit therefore give

$$\max_i \left| \|q_i\|_2^2 - \frac{1}{\alpha} \right| = o_{\mathbb{P}}(1), \qquad \max_i \left| q_i^\top \mu_i - s_\alpha \right| = o_{\mathbb{P}}(1). \tag{107}$$

Indeed, if $S_\mu := Q_\mu C_\mu^\top = U_\mu \Sigma_\mu U_\mu^\top$, conditional Haar concentration gives $\max_i |S_\mu[i,i] - \|C_\mu\|_*/k| = o_{\mathbb{P}}(1)$, while $\|C_\mu\|_*/k \to s_\alpha$ and $q_i^\top(\mu_i - \bar{\mu}) = S_\mu[i,i]$ with $\|\bar{\mu}\|_2 = o_{\mathbb{P}}(1)$. Since $s_\alpha = m_{1/2}(\alpha)/\sqrt{\alpha}$ and $\kappa_{\mathrm{sp}}(\alpha) = m_{1/2}(\alpha)$ for $\alpha > 1$, (107) yields

$$\max_i \left| \frac{q_i^\top \mu_i}{\|q_i\|_2 \|\mu_i\|_2} - \kappa_{\mathrm{sp}}(\alpha) \right| = o_{\mathbb{P}}(1). \tag{108}$$

We now transfer (108) from Q_μ to arbitrary exact Spectral minimizers. Fix $W \in \mathcal{M}_{\mathrm{sp}}^X$, set $a := \|W\|_{\mathrm{op}}$, and let $\bar{w} := k^{-1}\sum_i w_i$ and $h := \bar{w}/a$. Centering preserves every margin and cannot increase the operator norm, so exact optimality gives $\|PW\|_{\mathrm{op}} = \|W\|_{\mathrm{op}} = a$. Hence, with $B := PW/a$,

$$\frac{W}{a} = B + \mathbf{1}h^\top, \qquad \mathbf{1}^\top B = \mathbf{0}^\top. \tag{109}$$

Because $\|Q_\mu\|_{\mathrm{F}}^2 = d$, (106) gives $\|B\|_{\mathrm{F}}^2 = d - o_{\mathbb{P}}(d)$ uniformly over $\mathcal{M}_{\mathrm{sp}}^X$. The two terms in (109) are Frobenius-orthogonal, while $\|W/a\|_{\mathrm{op}} = 1$ implies $\|W/a\|_{\mathrm{F}}^2 \le d$. Therefore

$$d \ge \left\| \frac{W}{a} \right\|_{\mathrm{F}}^2 = \|B\|_{\mathrm{F}}^2 + k\|h\|_2^2,$$

and hence

$$\sup_{W \in \mathcal{M}_{\mathrm{sp}}^X} \|h\|_2 = o_{\mathbb{P}}(1). \tag{110}$$

Consequently,

$$\sup_{W \in \mathcal{M}_{\mathrm{sp}}^X} \frac{1}{k}\sum_{i=1}^k \left\| \frac{w_i}{a} - q_i \right\|_2^2 = \sup_{W \in \mathcal{M}_{\mathrm{sp}}^X} \left\{ \frac{1}{k}\|B - Q_\mu\|_{\mathrm{F}}^2 + \|h\|_2^2 \right\} = o_{\mathbb{P}}(1). \tag{111}$$

Moreover, every spectral row is nonzero with probability tending to one: if $w_i = \mathbf{0}$, then (56) gives $-ah^\top m_i \ge 1 - 1/k$, contradicting $a = O_{\mathbb{P}}(1)$, (110), and $\max_i \|m_i\|_2 = O_{\mathbb{P}}(1)$. Since (107) bounds $\min_i \|q_i\|_2$ away from zero, (111) and normalization stability imply that, for every fixed $\varepsilon > 0$,

$$\sup_{W \in \mathcal{M}_{\mathrm{sp}}^X} \frac{1}{k} \# \left\{ i : \left| \arccos\left(\frac{w_i^\top \mu_i}{\|w_i\|_2 \|\mu_i\|_2} \right) - \arccos \kappa_{\mathrm{sp}}(\alpha) \right| > \varepsilon \right\} \xrightarrow{\mathbb{P}} 0. \tag{112}$$

Because every angle lies in $[0, \pi]$, (112) implies

$$\sup_{W \in \mathcal{M}_{\mathrm{sp}}^X} \left| \frac{1}{k}\sum_{i=1}^k \arccos\left(\frac{w_i^\top \mu_i}{\|w_i\|_2 \|\mu_i\|_2} \right) - \arccos \kappa_{\mathrm{sp}}(\alpha) \right| \xrightarrow{\mathbb{P}} 0.$$

Finally, (37) identifies each maximum-margin direction set with the positive rays through the corresponding exact minimum-norm face. The full-batch implicit-bias results used in Theorem 3.1, together with Lemma B.2, give direction-set convergence to $\mathcal{U}_{\mathrm{row}}^X$ for RowGD and RowGDM, to $\mathcal{U}_{\mathrm{sp}}^X$ for SpecGD and Muon, and to $\mathcal{U}_{\max}^X$ for SignGD, Signum, and Adam. At each fixed d, the normalized row directions, the centered spectral representative $PW/\|PW\|_{\mathrm{op}}$, and the mean class–row angle are continuous on the corresponding compact direction sets on the high-probability events established above. Taking $t \to \infty$ first and then $d \to \infty$ therefore transfers (103), (105), (106), and the three angle limits to the optimizer trajectories, proving Proposition 3.2. □

C.9 Quantitative Population-error Exponents for Deterministic Geometries

Because all three accuracies may converge to one, a strict accuracy ordering can still correspond to an exponentially large separation in population error. To quantify this separation, define the directed-pair exponent and normalized log error by

$$\mathsf{E}_d(W) := \min_{i \ne j} \left[(u_{ij}^W)^\top \mu_i \right]_+^2, \qquad \mathsf{F}_d(W) := -\frac{2r_{\mathrm{te}}^2}{d} \log \mathcal{E}_{\pi,d}(W).$$

Both are invariant under positive rescaling and common row shifts.

Lemma C.24 (Directed-pair exponent controls multiclass error)**.** *For every classifier with nonzero pairwise row differences, set* $I_d(W) := d\mathsf{E}_d(W)/(2r_{\mathrm{te}}^2)$ *and* $C_0 := 3/2 + \frac{1}{2}\log(2\pi)$*. Then*

$$\max\{0, I_d(W) - \log(k-1)\} \le -\log \mathcal{E}_{\pi,d}(W) \le I_d(W) + \log(1/\pi_{\min,d}) + \log(1+\sqrt{2I_d(W)}) + C_0. \tag{113}$$

Consequently, uniformly over any classifier family on which $\mathsf{E}_d(W) = O(1)$*, one has* $\mathsf{F}_d(W) = \mathsf{E}_d(W) + o(1)$ *under Assumption 2.2.*

Proof. For a directed pair, the exact error is $\overline{\Phi}(\sqrt{d}\,(u_{ij}^W)^\top \mu_i / r_{\mathrm{te}})$. Let $t = \sqrt{2I_d(W)}$. The Gaussian tail bounds

$$\frac{e^{-C_0}}{1+t} e^{-t^2/2} \le \overline{\Phi}(t) \le \frac{1}{2} e^{-t^2/2}$$

show that one directed pair contributes at least $\pi_{\min,d} e^{-C_0 - I_d(W)}/(1+\sqrt{2I_d(W)})$ to the multiclass error, while a classwise union bound gives $\mathcal{E}_{\pi,d}(W) \le (k-1)e^{-I_d(W)}$ when all directed population margins are nonnegative; if one is negative then $I_d(W) = 0$, and the same logarithmic lower bound is trivial. This proves (113). Multiplying by $2r_{\mathrm{te}}^2/d$ and using $k = O(d)$, $\log(1/\pi_{\min,d}) = o(d)$, and $r_{\mathrm{te}} \le r_{\mathrm{te},0}$ proves the uniform $o(1)$ reduction. □

Let $g := \sqrt{2/\pi}$. The following sharpening is needed only for the quantitative statement; the qualitative accuracy theorem above already follows from the angular gaps in Theorem C.2.

Proposition C.25 (Sharp directed signals on the Max and Spectral faces)**.** *Under Assumptions 2.1 and 2.2, in probability,*

$$r^*_{\max}(X) = \frac{1+o(1)}{g\sqrt{d}}, \qquad r^*_{\mathrm{sp}}(X) = \frac{1+o(1)}{s_\alpha}. \tag{114}$$

Moreover, uniformly over the complete exact faces,

$$\sup_{W \in \mathcal{M}^X_{\max}} \max_{i \ne j} \left| (u_{ij}^W)^\top \mu_i - \frac{1}{\sqrt{\pi}} \right| = o_{\mathbb{P}}(1), \tag{115}$$

$$\inf_{W \in \mathcal{M}^X_{\mathrm{sp}}} \min_{i \ne j} (u_{ij}^W)^\top \mu_i \ge \frac{s_\alpha}{\sqrt{2}} - o_{\mathbb{P}}(1). \tag{116}$$

Proof. Let M_μ have rows $\mu_i^\top$, let $C_\mu := PM_\mu$, and let $T := \mathrm{sign}(M_\mu)$. Standard Gaussian concentration, uniformly over the $k = \Theta(d)$ classes, gives

$$\frac{\|C_\mu\|_{\mathrm{sum}}}{k\sqrt{d}} = g + o_{\mathbb{P}}(1), \qquad \max_i \left| \frac{\|\mu_i\|_1}{\sqrt{d}} - g \right| = o_{\mathbb{P}}(1), \qquad \max_{i \ne j} \frac{|T_j^\top \mu_i|}{\sqrt{d}} = o_{\mathbb{P}}(1), \tag{117}$$

and, for every fixed $a > 0$,

$$\frac{1}{kd} \#\{(i,\ell) : \sqrt{d}\,|C_\mu[i,\ell]| \le a\} \xrightarrow{\mathbb{P}} \mathbb{P}(|G| \le a), \qquad \max_i \frac{1}{d} \#\{\ell : \sqrt{d}\,|\mu_i[\ell]| \le a\} \le \mathbb{P}(|G| \le a) + o_{\mathbb{P}}(1). \tag{118}$$

Also $\|C_m - C_\mu\|_{\mathrm{sum}} = o_{\mathbb{P}}(k\sqrt{d})$ by Proposition C.4. Every feasible W satisfies the averaged exposure inequality $\langle W, C_m \rangle \ge k - 1$. Max–sum duality and (117) therefore give $r^*_{\max}(X) \ge (1 + o_{\mathbb{P}}(1))/(g\sqrt{d})$, while Proposition C.4, item (H5) gives the matching upper bound. Similarly, operator–nuclear duality and Proposition C.12, item (S1) give $r^*_{\mathrm{sp}}(X) \ge (1 + o_{\mathbb{P}}(1))/s_\alpha$, while the polar witness Q_m/ρ_X and Proposition C.12, item (S3) give the matching upper bound. This proves (114).

Fix $W \in \mathcal{M}^X_{\max}$ and set $A := W/r^*_{\max}(X)$, so $\|A\|_{\max} = 1$. The exposure inequality and the exact Max radius imply

$$\langle A, C_\mu \rangle \ge \|C_\mu\|_{\mathrm{sum}} - o_{\mathbb{P}}(k\sqrt{d})$$

uniformly over the complete Max face. Put $S := \mathrm{sign}(C_\mu)$ and $\Delta_{i\ell} := 1 - A[i,\ell]S[i,\ell] \in [0,2]$. Since $(A[i,\ell] - S[i,\ell])^2 = \Delta_{i\ell}^2 \le 2\Delta_{i\ell}$, for every fixed $a > 0$,

$$\frac{\|A - S\|_{\mathrm{F}}^2}{kd} \le \frac{2\sqrt{d}}{akd} \left(\|C_\mu\|_{\mathrm{sum}} - \langle A, C_\mu \rangle \right) + \frac{4}{kd} \#\{(i,\ell) : \sqrt{d}\,|C_\mu[i,\ell]| \le a\}.$$

First sending $d \to \infty$ and then $a \downarrow 0$ gives $\sup_W \|A - S\|_{\mathrm{F}}^2/(kd) = o_{\mathbb{P}}(1)$. Since $C_\mu[i,\ell] = \mu_i[\ell] - \bar{\mu}[\ell]$, where $\bar{\mu} = k^{-1}\sum_i \mu_i$ and $\|\bar{\mu}\|_2 = o_{\mathbb{P}}(1)$, the same small-ball argument transfers S to T, yielding

$$\sup_{W \in \mathcal{M}_{\max}^X} \frac{\|A - T\|_{\mathrm{F}}^2}{kd} = o_{\mathbb{P}}(1). \tag{119}$$

Let $\bar{A} = k^{-1}\sum_i A_i$ and $\bar{T} = k^{-1}\sum_i T_i$. Jensen's inequality and (119), together with $\mathbb{E}\|\bar{T}\|_2^2 = d/k = O(1)$, imply $\sup_W \|\bar{A}\|_2/\sqrt{d} = o_{\mathbb{P}}(1)$. Averaging the class-$i$ constraints over its samples and over all competitors gives

$$(A_i - \bar{A})^\top m_i \geq \frac{k-1}{k\, r_{\max}^*(X)} = g\sqrt{d} - o_{\mathbb{P}}(\sqrt{d}).$$

Proposition C.4, item (H1), the bound $\|A_i - \bar{A}\|_2 \leq 2\sqrt{d}$, and the preceding control of $\bar{A}$ therefore imply $A_i^\top \mu_i \geq g\sqrt{d} - o_{\mathbb{P}}(\sqrt{d})$ uniformly over i and the complete face. The nonnegative deficiency $D_i := \|\mu_i\|_1 - A_i^\top \mu_i$ is thus $o_{\mathbb{P}}(\sqrt{d})$ uniformly. For every fixed $a > 0$,

$$\frac{1}{d}\|A_i - T_i\|_2^2 \leq \frac{2D_i}{a\sqrt{d}} + \frac{4}{d}\#\{\ell : \sqrt{d}\,|\mu_i[\ell]| \leq a\}. \tag{120}$$

Using (118) and then sending $a \downarrow 0$ proves $\sup_W \max_i \|A_i - T_i\|_2/\sqrt{d} = o_{\mathbb{P}}(1)$. Finally, uniformly over $i \neq j$, $(T_i - T_j)^\top \mu_i = g\sqrt{d} + o_{\mathbb{P}}(\sqrt{d})$ and $\|T_i - T_j\|_2 = \sqrt{2d} + o_{\mathbb{P}}(\sqrt{d})$, which proves (115).

For Spectral, class-average feasibility gives $(w_i - w_j)^\top m_i \geq 1$, while $\|w_i - w_j\|_2 \leq \sqrt{2}\|W\|_{\mathrm{op}} = \sqrt{2}\, r_{\mathrm{sp}}^*(X)$. Proposition C.4, item (H1) and (114) therefore give $(u_{ij}^W)^\top \mu_i \geq s_\alpha/\sqrt{2} - o_{\mathbb{P}}(1)$ uniformly, proving (116). □

Proposition C.26 (Deterministic population-error exponents)**.** *Under Assumptions 2.1 and 2.2, uniformly over the indicated complete maximum-margin direction sets,*

$$\sup_{R \in \mathcal{U}_{\mathrm{row}}^X} \left| \mathsf{F}_d(R) - \frac{1}{2} \right| = o_{\mathbb{P}}(1),$$

$$\sup_{M \in \mathcal{U}_{\max}^X} \left| \mathsf{F}_d(M) - \frac{1}{\pi} \right| = o_{\mathbb{P}}(1),$$

$$\frac{s_\alpha^2}{2} - o_{\mathbb{P}}(1) \leq \mathsf{F}_d(S) \leq \frac{\kappa_{\mathrm{sp}}(\alpha)^2}{2} + o_{\mathbb{P}}(1), \qquad S \in \mathcal{U}_{\mathrm{sp}}^X.$$

Consequently, for $0 < \alpha \leq 1$, the spectral interval collapses to $\kappa_{\mathrm{sp}}(\alpha)^2/2$; at $\alpha = 1$, this value is $32/(9\pi^2)$. Uniformly over $R \in \mathcal{U}_{\mathrm{row}}^X$, $M \in \mathcal{U}_{\max}^X$, and $S \in \mathcal{U}_{\mathrm{sp}}^X$,

$$\log \frac{\mathcal{E}_{\pi,d}(M)}{\mathcal{E}_{\pi,d}(R)} = \left(\frac{\pi - 2}{4\pi} + o_{\mathbb{P}}(1) \right) \frac{d}{r_{\mathrm{te}}^2}, \tag{121}$$

$$\frac{1 - \kappa_{\mathrm{sp}}(\alpha)^2}{2} - o_{\mathbb{P}}(1) \leq \frac{2r_{\mathrm{te}}^2}{d} \log \frac{\mathcal{E}_{\pi,d}(S)}{\mathcal{E}_{\pi,d}(R)} \leq \frac{1 - s_\alpha^2}{2} + o_{\mathbb{P}}(1). \tag{122}$$

The Row–Spectral gain is exact when $\alpha \leq 1$.

Proof. For Row, Proposition C.7 and the Gaussian mean shell imply $(u_{ij}^R)^\top \mu_i = 1/\sqrt{2} + o_{\mathbb{P}}(1)$ uniformly over all directed pairs and the complete Row face. Proposition C.25 gives the corresponding uniform Max signal $1/\sqrt{\pi} + o_{\mathbb{P}}(1)$. Hence $\mathsf{E}_d(R) = 1/2 + o_{\mathbb{P}}(1)$ and $\mathsf{E}_d(M) = 1/\pi + o_{\mathbb{P}}(1)$ uniformly.

For Spectral, Proposition C.25 gives $\mathsf{E}_d(S) \geq s_\alpha^2/2 - o_{\mathbb{P}}(1)$ uniformly. Theorem C.2 implies that every spectral direction has an ordered pair with $d_{ij}(S) \geq \delta_{\mathrm{sp}}(\alpha) - o_{\mathbb{P}}(1)$, and hence $(u_{ij}^S)^\top u_{ij}^\mu \leq \kappa_{\mathrm{sp}}(\alpha) + o_{\mathbb{P}}(1)$. For that unordered pair,

$$(u_{ij}^S)^\top \mu_i + (u_{ji}^S)^\top \mu_j = (u_{ij}^S)^\top (\mu_i - \mu_j) \leq \sqrt{2}\,\kappa_{\mathrm{sp}}(\alpha) + o_{\mathbb{P}}(1).$$

Both directed signals are positive by the preceding lower bound, so at least one is at most $\kappa_{\mathrm{sp}}(\alpha)/\sqrt{2} + o_{\mathbb{P}}(1)$. This proves the upper endpoint for $\mathsf{E}_d(S)$. Lemma C.24 transfers all three statements from E_d to F_d. Finally, $s_\alpha = \kappa_{\mathrm{sp}}(\alpha)$ for $\alpha \leq 1$, while subtraction of the exponents gives (121)–(122). □

Proposition C.27 (Population-error exponents of deterministic optimizers)**.** *Under the assumptions and optimizer conventions of Theorem 3.1, the deterministic full-batch optimizer trajectories satisfy, in probability,*

$$\limsup_{t\to\infty}\left|\mathsf{F}_d\big(W_t^{\mathrm{RowGDM}}\big)-\frac{1}{2}\right|=o_{\mathbb{P}}(1),\qquad \limsup_{t\to\infty}\left|\mathsf{F}_d\big(W_t^{\mathrm{Adam}}\big)-\frac{1}{\pi}\right|=o_{\mathbb{P}}(1). \tag{123}$$

Exact-SVD Muon satisfies

$$\frac{s_\alpha^2}{2}-o_{\mathbb{P}}(1)\le\liminf_{t\to\infty}\mathsf{F}_d\big(W_t^{\mathrm{Muon}}\big)\le\limsup_{t\to\infty}\mathsf{F}_d\big(W_t^{\mathrm{Muon}}\big)\le\frac{\kappa_{\mathrm{sp}}(\alpha)^2}{2}+o_{\mathbb{P}}(1). \tag{124}$$

Define the Row–Spectral exponent-gap constant

$$\Delta_{\exp}(\alpha):=\frac{1-\kappa_{\mathrm{sp}}(\alpha)^2}{4}=\frac{1}{4}\left(1-\frac{m_{1/2}(\alpha)^2}{\alpha\wedge 1}\right)>0,\qquad m_{1/2}(\alpha):=\frac{\alpha}{2\pi}\int_{a_\alpha}^{b_\alpha}\frac{\sqrt{(b_\alpha-x)(x-a_\alpha)}}{\sqrt{x}}\,dx, \tag{125}$$

where

$$a_\alpha=(1-\alpha^{-1/2})^2,\qquad b_\alpha=(1+\alpha^{-1/2})^2.$$

Equivalently,

$$\Delta_{\exp}(\alpha)=\frac{1}{4}\left[1-\frac{\alpha^2}{4\pi^2(\alpha\wedge 1)}\left(\int_{a_\alpha}^{b_\alpha}\frac{\sqrt{(b_\alpha-x)(x-a_\alpha)}}{\sqrt{x}}\,dx\right)^2\right].$$

For example,

$$\Delta_{\exp}(1)=\frac{1}{4}\left(1-\frac{64}{9\pi^2}\right)\approx 0.06987,\qquad \Delta_{\exp}(2)\approx 0.03247,\qquad \Delta_{\exp}(3)\approx 0.02133.$$

Consequently,

$$\limsup_{t\to\infty}\left|\frac{r_{\mathrm{te}}^2}{d}\log\frac{\mathcal{E}_{\pi,d}(W_t^{\mathrm{Adam}})}{\mathcal{E}_{\pi,d}(W_t^{\mathrm{RowGDM}})}-\frac{\pi-2}{4\pi}\right|=o_{\mathbb{P}}(1),$$

and

$$\begin{aligned}\Delta_{\exp}(\alpha)-o_{\mathbb{P}}(1)&\le\liminf_{t\to\infty}\frac{r_{\mathrm{te}}^2}{d}\log\frac{\mathcal{E}_{\pi,d}(W_t^{\mathrm{Muon}})}{\mathcal{E}_{\pi,d}(W_t^{\mathrm{RowGDM}})}\\&\le\limsup_{t\to\infty}\frac{r_{\mathrm{te}}^2}{d}\log\frac{\mathcal{E}_{\pi,d}(W_t^{\mathrm{Muon}})}{\mathcal{E}_{\pi,d}(W_t^{\mathrm{RowGDM}})}\le\frac{1-s_\alpha^2}{4}+o_{\mathbb{P}}(1).\end{aligned}$$

The same conclusions hold with RowGDM replaced by RowGD, with Adam replaced by SignGD or Signum, and with Muon replaced by Spectral-GD. When $0<\alpha\le 1$, the two Row–Muon bounds coincide. Here, $o_{\mathbb{P}}(1)$ refers to the limit $d\to\infty$, after taking $t\to\infty$ at each fixed d.

Proof. The full-batch margin-convergence results used in Theorem 3.1 imply that the normalized trajectories approach the corresponding row-, entrywise max-, and spectral-norm maximum-margin direction sets. Since Proposition C.26 holds uniformly over each complete direction set, Lemma C.21 transfers its exponent bounds to the long-time optimizer trajectories. This gives (123)–(124). The definition (125) follows from $\kappa_{\mathrm{sp}}(\alpha)=m_{1/2}(\alpha)/\sqrt{\alpha\wedge 1}$, and the integral form follows from the Marchenko–Pastur density. The log-error ratio statements follow from

$$\frac{r_{\mathrm{te}}^2}{d}\log\frac{\mathcal{E}_{\pi,d}(W^{(2)})}{\mathcal{E}_{\pi,d}(W^{(1)})}=\frac{1}{2}\left(\mathsf{F}_d(W^{(1)})-\mathsf{F}_d(W^{(2)})\right).$$

No uniqueness of the optimizer-selected limiting direction is required. □

Corollary C.28 (Full-batch Muon versus Adam)**.** *Under the assumptions and optimizer conventions of Theorem 3.1, suppose $0 < \alpha < \alpha_\star$, where $\alpha_\star \approx 1.225$ is the aspect-ratio threshold defined by $s^2_{\alpha_\star} = 2/\pi$. Then, with*

$$c_{\mathrm{MA}}(\alpha) := \frac{1}{4}\left(s_\alpha^2 - \frac{2}{\pi}\right) > 0,$$

the deterministic full-batch trajectories satisfy

$$\limsup_{t\to\infty} \frac{r_{\mathrm{te}}^2}{d} \log \frac{\mathcal{E}_{\pi,d}(W_t^{\mathrm{Muon}})}{\mathcal{E}_{\pi,d}(W_t^{\mathrm{Adam}})} \le -c_{\mathrm{MA}}(\alpha) + o_{\mathbb{P}}(1). \tag{126}$$

Thus exact-SVD Muon has asymptotically smaller population error, and hence higher population accuracy, than full-batch Adam with probability tending to one. The same conclusion holds with Muon replaced by Spectral-GD and Adam replaced by SignGD or Signum.

Proof. Proposition C.27 gives $\liminf_{t\to\infty} \mathsf{F}_d(W_t^{\mathrm{Muon}}) \ge s_\alpha^2/2 - o_{\mathbb{P}}(1)$ and $\mathsf{F}_d(W_t^{\mathrm{Adam}}) = 1/\pi + o_{\mathbb{P}}(1)$ at long times. Using $\frac{r_{\mathrm{te}}^2}{d}\log(\mathcal{E}_2/\mathcal{E}_1) = \frac{1}{2}(\mathsf{F}_d(W_1) - \mathsf{F}_d(W_2))$ gives (126). □

Remark C.29 (Relation to last-layer practice)**.** *In practical last-layer training, Adam is typically preferred to Muon, whereas Corollary C.28 predicts the opposite ordering for $\alpha < \alpha_\star$. A likely source of this mismatch is that our result compares only asymptotic convergence directions and therefore does not capture convergence speed or finite-time trajectories. The comparison is also idealized, using deterministic full-batch training, exact-SVD Muon, and Adam without a numerical-stability constant. For $\alpha \ge \alpha_\star$, the ordering between Adam and Muon remain open.*

D Proofs for Stochastic Optimizers

This appendix proves Proposition 3.4. The random-reshuffling momentum analysis in Corollary B.12 places RR-RowSGDM, RR-Muon, and RR-Signum in near-maximum-margin direction sets for the row, spectral, and entrywise max norms, respectively. We show that, for sufficiently large momentum, these near-margin sets retain enough of the deterministic geometric separation to yield a strict population-accuracy ordering. Throughout this appendix, training time tends to infinity at each fixed dimension before the limit $d \to \infty$.

For each d, write $m_d := N(d)/b_d$. Recall the geometry-adapted atom radius $R_\psi(X)$ from (17), and for $\psi \in \{\psi_{\mathrm{row}}, \psi_{\mathrm{sp}}, \psi_{\max}\}$ define

$$\varepsilon_{\psi,\beta,b}(X) := 2(1-\beta)m_d(m_d^2 - 1)R_\psi(X), \qquad \mathcal{U}^X_{\psi,\beta,b} := \mathcal{U}^X_{\psi,\varepsilon_{\psi,\beta,b}(X)}.$$

D.1 Near-margin directions and near-minimum-norm representatives

Lemma D.1 (Near-margin reduction)**.** *Fix a linearly separable dataset X, a matrix norm ψ, and $0 \le \varepsilon < \gamma_\psi(X)$. For every $U \in \mathcal{U}^X_{\psi,\varepsilon}$, define*

$$V_U := \frac{U}{\mathrm{mar}_X(U)}, \qquad \xi_\psi(\varepsilon; X) := \frac{\varepsilon}{\gamma_\psi(X) - \varepsilon}.$$

Then $V_U \in \mathcal{F}_X$,

$$\psi(V_U) \le (1 + \xi_\psi(\varepsilon; X))\, r^*_\psi(X),$$

and V_U has the same normalized pairwise row differences as U.

Proof. Set $\widehat{U} := U/\psi(U)$. By definition of $\mathcal{U}^X_{\psi,\varepsilon}$, $\mathrm{mar}_X(\widehat{U}) \ge \gamma_\psi(X) - \varepsilon > 0$. Hence $V_U = \widehat{U}/\mathrm{mar}_X(\widehat{U})$ is feasible and

$$\psi(V_U) = \frac{1}{\mathrm{mar}_X(\widehat{U})} \le \frac{1}{\gamma_\psi(X) - \varepsilon} = \frac{r^*_\psi(X)}{1 - \varepsilon/\gamma_\psi(X)} = (1 + \xi_\psi(\varepsilon; X))\, r^*_\psi(X).$$

The normalized pairwise row differences are unchanged under positive scaling. □

D.2 Quantitative stability of near minimizers

Lemma D.2 (Combined Gaussian-cloud event)**.** *There are deterministic $\omega_d \downarrow 0$, constants $C, c > 0$, and events $\mathcal{E}_d^{\mathrm{near}}$ such that*

$$\mathbb{P}((\mathcal{E}_d^{\mathrm{near}})^c) \le C \exp\left[-c\sqrt{d \log(2N(d)(d+1))}\right], \tag{127}$$

all conclusions of Propositions C.4 and C.12 hold with their vanishing errors bounded by ω_d, and

$$\max_{i,s} \|x_{i,s}\|_2 \le 1 + 2r_0 + \omega_d, \qquad \max_{i,s} \|x_{i,s}\|_1 \le (1 + 2r_0 + \omega_d)\sqrt{d}. \tag{128}$$

We may additionally take $\omega_d \ge 1/k$.

Proof. Intersect the events in Propositions C.4 and C.12 with the event $\max_{i,s} \|z_{i,s}\|_2 \le 2$. The latter fails with probability at most $2Ne^{-cd}$, which is smaller than the right-hand side of (127). Under the normalized convention (38), the mean shell gives $\max_i \|\mu_i\|_2 \le 1 + \omega_d$, so $\|x_{i,s}\|_2 \le 1 + \omega_d + 2r_0$, and the ℓ_1 bound follows from $\|x\|_1 \le \sqrt{d}\,\|x\|_2$. Enlarge one deterministic envelope to dominate all finitely many errors and $1/k$. □

Proposition D.3 (Near-minimizer stability)**.** *There are constants $C_{\mathrm{row}}, C_{\max} > 0$ and $C_{\mathrm{sp}}(\alpha) > 0$ such that, on $\mathcal{E}_d^{\mathrm{near}}$, for every $0 \le \xi \le 1$,*

$$W \in \mathcal{F}_X,\ \|W\|_{\mathrm{row}} \le (1+\xi) r^*_{\mathrm{row}}(X) \Longrightarrow \max_{i \ne j} \|u_{ij}^W - u_{ij}^\mu\|_2 \le C_{\mathrm{row}}(\omega_d + \xi)^{1/4}, \tag{129}$$

$$W \in \mathcal{F}_X,\ \|W\|_{\mathrm{op}} \le (1+\xi) r^*_{\mathrm{sp}}(X) \Longrightarrow D_{\text{avg-bdry}}(W) \ge \delta_{\mathrm{sp}}(\alpha) - C_{\mathrm{sp}}(\alpha)(\omega_d + \xi)^{1/4}, \tag{130}$$

$$W \in \mathcal{F}_X,\ \|W\|_{\max} \le (1+\xi) r^*_{\max}(X) \Longrightarrow D_{\text{avg-bdry}}(W) \ge \delta_{\max} - C_{\max}(\omega_d + \xi). \tag{131}$$

Proof. For the row norm, put $q := \omega_d + \xi$. The empirical-mean witness in Proposition C.6 gives $r^*_{\mathrm{row}}(X) \le 1 + C\omega_d$, so $\max_i \|w_i\|_2 \le 1 + Cq$. Averaging feasibility within class i and then over all competing classes gives $(w_i - \bar{w})^\top m_i \ge 1 - 1/k$. The row variance identity then gives $\|\bar{w}\|_2^2 \le Cq$, and substituting this bound back into the averaged margin inequality yields $\max_i \|w_i - m_i\|_2 \le Cq^{1/4}$. The mean perturbation in Proposition C.4 and Lemma C.5 prove (129).

For the spectral norm, Proposition C.15 gives an error $C_\alpha q^{1/2}$ when $\alpha \ne 1$ and $Cq^{1/4}$ when $\alpha = 1$. Since $q^{1/2} \le q^{1/4}$ for $q \le 1$, enlarging the constant proves (130) uniformly over all aspect ratios.

For the max norm, fix $i \ne j$, set $v := w_i - w_j$, and put $u := u_{ij}^m$. Near-optimality and Proposition C.4(H5) give

$$\|v\|_\infty \le \frac{\sqrt{2\pi} + Cq}{\sqrt{d}}.$$

The two class-averaged feasibility constraints imply $v^\top(m_i - m_j) \ge 2$, and hence $v^\top u \ge \sqrt{2} - Cq$. Applying the clipped minimum-energy characterization in Proposition C.8, the uniform empirical laws in Proposition C.4(H6), and Lemma C.9 gives

$$\frac{v^\top u}{\|v\|_2} \le \kappa_{\max} + Cq.$$

The cosine-distance identity and Proposition C.4(H4) prove (131). □

D.3 A dimension-independent momentum threshold

Proposition D.4 (Uniform radius-to-margin ratios)**.** *On $\mathcal{E}_d^{\mathrm{near}}$, for every sufficiently large d,*

$$\frac{R_{\mathrm{row}}(X)}{\gamma_{\mathrm{row}}(X)} \le K_{\mathrm{row}}, \qquad \frac{R_{\mathrm{sp}}(X)}{\gamma_{\mathrm{sp}}(X)} \le K_{\mathrm{sp}}, \qquad \frac{R_{\max}(X)}{\gamma_{\max}(X)} \le K_{\max}, \tag{132}$$

where one may take

$$K_{\mathrm{row}} := 4(1 + 2r_0), \qquad K_{\mathrm{sp}} := \frac{8(1 + 2r_0)}{s_\alpha}, \qquad K_{\max} := (\sqrt{2\pi} + 1)(1 + 2r_0).$$

Proof. Each ratio equals $R_\psi(X)r^*_\psi(X)$. For every sufficiently large d, Proposition C.6 gives $r^*_{\mathrm{row}}(X) \le 2$, the reduced-polar witness (89) and $\rho_X \to s_\alpha > 0$ give $r^*_{\mathrm{sp}}(X) \le 4/s_\alpha$, and Proposition C.4(H5) gives $r^*_{\max}(X) \le (\sqrt{2\pi}+1)/(2\sqrt{d})$. Equation (128) and $\omega_d \to 0$ imply, for sufficiently large d,

$$R_{\mathrm{row}}(X) = R_{\mathrm{sp}}(X) \le 2(1+2r_0), \qquad R_{\max}(X) \le 2(1+2r_0)\sqrt{d}.$$

Combining these estimates proves (132). □

Let

$$\delta_* := \min\{\delta_{\mathrm{sp}}(\alpha), \delta_{\max}\} > 0. \tag{133}$$

Choose $h_* > 0$ sufficiently small that

$$\frac{1}{\sqrt{2}} - h_* > \frac{1}{\sqrt{2}}\left(1 - \frac{(\delta_*/2)^2}{2}\right)_+. \tag{134}$$

Choose $b_* \in (0,1]$ sufficiently small that

$$C_{\mathrm{row}} b_*^{1/4} \le h_*, \qquad C_{\mathrm{sp}}(\alpha) b_*^{1/4} \le \frac{\delta_*}{8}, \qquad C_{\max} b_* \le \frac{\delta_*}{8}.$$

Set

$$\theta_* := \frac{b_*/2}{1+b_*/2}, \qquad A_0 := m_0(m_0^2-1), \qquad K_* := \max\{K_{\mathrm{row}}, K_{\mathrm{sp}}, K_{\max}\},$$

and define

$$\beta_0 := \begin{cases} \dfrac{1}{2}, & A_0 = 0, \\ \max\left\{\dfrac{1}{2}, 1 - \dfrac{\theta_*}{2A_0K_*}\right\}, & A_0 > 0. \end{cases} \tag{135}$$

All constants in (133)–(135) depend only on (α, r_0, m_0).

Proposition D.5 (Uniform stochastic near-margin geometry)**.** *Under the data and batching assumptions of Proposition 3.4, fix $\beta \in [\beta_0, 1)$. Then, for every sufficiently large d, on $\mathcal{E}_d^{\mathrm{near}}$, all three effective margins $\gamma_\psi(X) - \varepsilon_{\psi,\beta,b}(X)$ are positive and*

$$\sup_{U \in \mathcal{U}^X_{\mathrm{row},\beta,b}} \max_{i \ne j} \|u_{ij}^U - u_{ij}^\mu\|_2 \le h_*, \tag{136}$$

$$\inf_{V \in \mathcal{U}^X_{\mathrm{sp},\beta,b}} D_{\text{avg-bdry}}(V) \ge \frac{\delta_*}{2},$$

$$\inf_{V \in \mathcal{U}^X_{\max,\beta,b}} D_{\text{avg-bdry}}(V) \ge \frac{\delta_*}{2}. \tag{137}$$

Moreover, after increasing the deterministic dimension threshold if necessary,

$$\mathbb{P}(\mathcal{E}_d^{\mathrm{near}}) \ge 1 - e^{-\sqrt{d}}.$$

Proof. For each geometry, define the relative stochastic deficit

$$\Delta_{\psi,d}(\beta) := \frac{\varepsilon_{\psi,\beta,b}(X)}{\gamma_\psi(X)} = 2(1-\beta)m_d(m_d^2-1)\frac{R_\psi(X)}{\gamma_\psi(X)}.$$

Since $m_d \le m_0$, Proposition D.4 gives

$$\Delta_{\psi,d}(\beta) \le 2(1-\beta)A_0K_* \le \theta_* < 1.$$

Thus every effective margin is positive. By the definition of θ_*,

$$\frac{\Delta_{\psi,d}(\beta)}{1-\Delta_{\psi,d}(\beta)} \le \frac{b_*}{2}.$$

For every $U \in \mathcal{U}^X_{\psi,\beta,b}$, Lemma D.1 therefore gives a feasible representative with identical normalized pairwise row differences and relative norm slack at most $b_*/2$. For every sufficiently large d, $\omega_d \le b_*/2$, so Proposition D.3 gives

$$\sup_{U\in\mathcal{U}^X_{\mathrm{row},\beta,b}} \max_{i\neq j} \|u^U_{ij} - u^\mu_{ij}\|_2 \le C_{\mathrm{row}} b_*^{1/4} \le h_*,$$

$$\inf_{V\in\mathcal{U}^X_{\mathrm{sp},\beta,b}} D_{\text{avg-bdry}}(V) \ge \delta_{\mathrm{sp}}(\alpha) - C_{\mathrm{sp}}(\alpha) b_*^{1/4} \ge \delta_* - \frac{\delta_*}{8} \ge \frac{\delta_*}{2},$$

and

$$\inf_{V\in\mathcal{U}^X_{\max,\beta,b}} D_{\text{avg-bdry}}(V) \ge \delta_{\max} - C_{\max} b_* \ge \delta_* - \frac{\delta_*}{8} \ge \frac{\delta_*}{2}.$$

This proves (136)–(137). Finally, $\log(2N(d)(d+1)) \to \infty$, so the failure probability in (127) is at most $e^{-\sqrt{d}}$ for every sufficiently large d. □

D.4 Proof of the stochastic optimizer accuracy ordering

Proof of Proposition 3.4. Fix $\beta \in [\beta_0, 1)$, and let $\mathcal{R}_d$ be the Gaussian radial event from Lemma C.18. Set

$$\mathcal{H}_d := \mathcal{E}^{\mathrm{near}}_d \cap \mathcal{R}_d.$$

By (127) and (95), after increasing the deterministic dimension threshold,

$$\mathbb{P}(\mathcal{H}_d) \ge 1 - e^{-\sqrt{d}}. \tag{138}$$

Work on $\mathcal{H}_d$. Define the Row family

$$\mathcal{R}^{\mathrm{clf}}_d := \mathcal{U}^X_{\mathrm{row},\beta,b}$$

and, for $\psi \in \{\mathrm{sp}, \max\}$, the competitor family

$$\mathcal{C}^{\mathrm{clf}}_{\psi,d} := \mathcal{U}^X_{\psi,\beta,b}.$$

Proposition D.5 gives

$$\sup_{U\in\mathcal{R}^{\mathrm{clf}}_d} \max_{i\neq j} \|u^U_{ij} - u^\mu_{ij}\|_2 \le h_*, \qquad \inf_{V\in\mathcal{C}^{\mathrm{clf}}_{\psi,d}} D_{\text{avg-bdry}}(V) \ge \frac{\delta_*}{2}.$$

The choice of h_* in (134) gives the strict limiting amplitude gap

$$\mathsf{a}(h_*) > \mathsf{b}(\delta_*/2),$$

where a and b are defined in (98).

Choose fixed constants $0 < b < a$ such that

$$\mathsf{b}(\delta_*/2) < b < a < \mathsf{a}(h_*).$$

On $\mathcal{R}_d$, the radial deviation and coherence in (94) converge to zero under a deterministic envelope. Therefore, for every sufficiently large d, the finite-dimensional amplitudes in (96)–(97) satisfy

$$A_d \ge a, \qquad B_d \le b$$

for the Row family and each of the two competitor families. Proposition C.19 consequently gives

$$\sup_{U\in\mathcal{R}^{\mathrm{clf}}_d} \mathcal{E}_{\pi,d}(U) \le (k-1)\overline{\Phi}\left(\frac{a}{\sigma_d}\right), \qquad \inf_{V\in\mathcal{C}^{\mathrm{clf}}_{\psi,d}} \mathcal{E}_{\pi,d}(V) \ge 2\pi_{\min,d}\overline{\Phi}\left(\frac{b}{\sigma_d}\right).$$

For $x \ge y \ge 0$, the Gaussian upper tail satisfies $\overline{\Phi}(x)/\overline{\Phi}(y) \le \exp(-(x^2-y^2)/2)$. Hence $k = O(d)$, $r_{\mathrm{te}} \le r_{\mathrm{te},0}$, and $\log(1/\pi_{\min,d}) = o(d)$ imply

$$\frac{\sup_{U\in\mathcal{R}^{\mathrm{clf}}_d} \mathcal{E}_{\pi,d}(U)}{\inf_{V\in\mathcal{C}^{\mathrm{clf}}_{\psi,d}} \mathcal{E}_{\pi,d}(V)} \le \frac{Ck}{\pi_{\min,d}} e^{-cd} \longrightarrow 0.$$

Hence, for every sufficiently large d on $\mathcal{H}_d$,

$$\sup_{U\in\mathcal{U}^X_{\mathrm{row},\beta,b}} \mathcal{E}_{\pi,d}(U) < \min_{\psi\in\{\mathrm{sp},\max\}} \inf_{V\in\mathcal{U}^X_{\psi,\beta,b}} \mathcal{E}_{\pi,d}(V). \tag{139}$$

It remains to transfer the set ordering to the optimizer trajectories. Proposition D.5 ensures that all effective margins are positive. Corollary B.12 therefore gives

$$\begin{aligned} d_{\cos}\big(W_t^{\text{RR-RowSGDM}}, \mathcal{U}^X_{\mathrm{row},\beta,b}\big) &\longrightarrow 0, \\ d_{\cos}\big(W_t^{\text{RR-Muon}}, \mathcal{U}^X_{\mathrm{sp},\beta,b}\big) &\longrightarrow 0, \\ d_{\cos}\Big(W_t^{\text{RR-Signum}}, \mathcal{U}^X_{\max,\beta,b}\Big) &\longrightarrow 0. \end{aligned} \tag{140}$$

Applying Lemma C.21 to (140) and then using (139) gives, for each $\mathrm{opt} \in \{\mathrm{Muon}, \mathrm{Signum}\}$,

$$\begin{aligned} \limsup_{t\to\infty} \mathcal{E}_{\pi,d}\big(W_t^{\text{RR-RowSGDM}}\big) &\le \sup_{U\in\mathcal{U}^X_{\mathrm{row},\beta,b}} \mathcal{E}_{\pi,d}(U) \\ &< \inf_{V\in\mathcal{U}^X_{\psi_{\mathrm{opt}},\beta,b}} \mathcal{E}_{\pi,d}(V) \\ &\le \liminf_{t\to\infty} \mathcal{E}_{\pi,d}\big(W_t^{\mathrm{opt}}\big), \end{aligned}$$

where $\psi_{\mathrm{Muon}} = \psi_{\mathrm{sp}}$ and $\psi_{\mathrm{Signum}} = \psi_{\max}$. Since $\mathrm{Acc}_{\pi,d} = 1 - \mathcal{E}_{\pi,d}$,

$$\liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\text{RR-RowSGDM}}\big) > \max_{\mathrm{opt}\in\{\mathrm{Muon},\mathrm{Signum}\}} \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big).$$

Since H_d and the dimension threshold above are independent of $\beta \in [\beta_0, 1)$, and the random-reshuffling margin bounds hold deterministically for every reshuffling realization, the conclusion holds simultaneously for all such β and every realization of the reshuffling. Together with (138), this proves (6). □

E Analysis of AdamProxy for Random-reshuffling Mini-batch Adam

This appendix proves Proposition 3.5, Proposition 3.6, and Proposition E.10. We first show that $\beta_1, \beta_2 \to 1$ random-reshuffling Adam is approximated over one epoch by the deterministic AdamProxy direction. We then prove a trajectory-level coordinate bound for AdamProxy and use it to establish a coordinate-induced angular gap, which, together with the near-row geometry from Appendix D and the geometry-to-accuracy transfer in Appendix C, yields the RR-RowSGDM–AdamProxy population-accuracy ordering. Separately, we characterize the limiting AdamProxy classifier through a self-consistent semidefinite quadratic margin problem when the normalized proxy direction converges.

E.1 One-epoch $\beta_1, \beta_2 \to 1$ RR Adam approximation

Throughout this subsection, probability is taken only over the independent random reshufflings while the training data are fixed. We assume $1 \le b < N$, use the gradient convention of the main text, namely $g_n(W) = -(e_{y_n} - p_n(W))x_n^\top$, and use the standard Adam initialization $M_0 = V_0 = \mathbf{0}$. For a uniformly random b-subset $B \subset [N]$, recall $G_B(W) = b^{-1}\sum_{n\in B} g_n(W)$, and set

$$S(W) := \sum_{n=1}^N g_n(W), \qquad Q(W) := \sum_{n=1}^N g_n(W)^{\odot 2}.$$

Lemma E.1 (Finite-population batch moments and relative scale)**.** *Assume $1 \le b < N$. For $z = (z_1, \dots, z_N) \in \mathbb{R}^N$, define $T_B z := b^{-1}\sum_{n\in B} z_n$ and $\sigma(z)^2 := \mathbb{E}_B[(T_B z)^2]$. Then*

$$\mathbb{E}_B[T_B z] = \frac{1}{N}\sum_{n=1}^N z_n, \tag{141}$$

$$\sigma(z)^2 = \frac{b-1}{bN(N-1)}\left(\sum_{n=1}^{N} z_n\right)^2 + \frac{N-b}{bN(N-1)}\sum_{n=1}^{N} z_n^2. \tag{142}$$

Consequently, with $\mu_1(W) := \mathbb{E}_B[G_B(W)]$ *and* $\mu_2(W) := \mathbb{E}_B[G_B(W)^{\odot 2}]$,

$$\mu_1(W) = \frac{1}{N}S(W), \qquad \mu_2(W) = \frac{(b-1)S(W)^{\odot 2} + (N-b)Q(W)}{bN(N-1)}, \tag{143}$$

and therefore

$$\mu_1(W) \oslash \mu_2(W)^{\odot 1/2} = \Phi_{N,b}(W). \tag{144}$$

Moreover, because $1 \le b < N$, σ *is a norm and*

$$\underline{c}_{N,b}\|z\|_2 \le \sigma(z) \le \frac{1}{\sqrt{b}}\|z\|_2, \qquad \underline{c}_{N,b} := \sqrt{\frac{N-b}{bN(N-1)}}, \tag{145}$$

while every b*-subset satisfies*

$$|T_B z| \le K_{N,b}\sigma(z), \qquad \frac{1}{b}\sum_{n\in B}|z_n| \le K_{N,b}\sigma(z), \qquad K_{N,b} := \sqrt{\frac{N(N-1)}{N-b}}. \tag{146}$$

For a matrix coordinate $\rho = (r, a)$, *define* $z_n^\rho(W) := g_n(W)[\rho]$ *and* $\sigma_\rho(W) := \sigma(z^\rho(W))$. *Under the nonzero-coordinate assumption in Proposition 3.5,* $\sigma_\rho(W) > 0$ *for every finite* W *and every* ρ.

Proof. Let $I_n := \mathbf{1}_{\{n\in B\}}$. Since B is a uniform b-subset, $\mathbb{E}I_n = b/N$ and $\mathbb{E}[I_n I_{n'}] = b(b-1)/(N(N-1))$ for $n \ne n'$. Expanding $\mathbb{E}[(T_B z)^2]$ gives (142), while (141) is immediate. Applying the two identities coordinate-wise to $g_n(W)$ gives (143), and substituting these expressions into $\mu_1(W) \oslash \mu_2(W)^{\odot 1/2}$ gives (144) by (7). The second term in (142) gives the lower bound in (145), while $|T_B z| \le b^{-1}\sum_{n\in B}|z_n| \le b^{-1/2}\|z\|_2$ gives the upper bound and (146). Finally, at finite W, every softmax probability is strictly positive. If $r = y_n$, then $g_n(W)[r,a] = -(1-p_{n,y_n}(W))x_n[a] \ne 0$, while if $r \ne y_n$, then $g_n(W)[r,a] = p_{n,r}(W)x_n[a] \ne 0$. Hence $z^\rho(W) \ne 0$, so $\sigma_\rho(W) > 0$. □

Lemma E.2 (Adam-step and signal stability)**.** *Let* $0 < \beta_1 \le \beta_2 < 1$ *and define*

$$A(\beta_1,\beta_2)^2 := \frac{(1-\beta_1)^2}{(1-\beta_2)(1-\beta_1^2/\beta_2)}.$$

Then $\beta_1^2 < \beta_2$, *and for* $U_t := M_{t+1} \oslash V_{t+1}^{\odot 1/2}$,

$$\|U_t\|_{\max} \le A(\beta_1,\beta_2), \qquad \|W_{t+1} - W_t\|_{\mathrm{F}} \le \sqrt{kd}\, A(\beta_1,\beta_2)\eta_t. \tag{147}$$

There are finite data-dependent constants L, C *such that, for all* U, W, *every batch* B, *and every coordinate* ρ,

$$\frac{|G_B(U)[\rho] - G_B(W)[\rho]|}{\sigma_\rho(W)} \le C\left(e^{L\|U-W\|_{\mathrm{F}}} - 1\right), \tag{148}$$

$$\frac{|G_B(U)[\rho]^2 - G_B(W)[\rho]^2|}{\sigma_\rho(W)^2} \le C\left(e^{2L\|U-W\|_{\mathrm{F}}} - 1\right). \tag{149}$$

Proof. Fix one coordinate and write $Y_u := G_{B_u}(W_u)[\rho]$. Zero initialization gives $M_{t+1} = (1-\beta_1)\sum_{j=0}^{t}\beta_1^j Y_{t-j}$ and $V_{t+1} = (1-\beta_2)\sum_{j=0}^{t}\beta_2^j Y_{t-j}^2$. If $V_{t+1} = 0$, then $M_{t+1} = 0$; otherwise Cauchy–Schwarz gives $M_{t+1}^2 \le A(\beta_1,\beta_2)^2 V_{t+1}$, proving (147). For the stability bounds, let $\Delta := U - W$. The logit perturbations satisfy $|e_c^\top \Delta x_n| \le \|\Delta\|_{\mathrm{F}}\|x_n\|_2$, so every softmax probability ratio, and also the ratio of $1-p_{n,y_n}$, lies between $e^{-L\|\Delta\|_{\mathrm{F}}}$ and $e^{L\|\Delta\|_{\mathrm{F}}}$ after enlarging L. Each coordinate of g_n is a fixed feature coordinate multiplied by one of these quantities, giving $|g_n(U)[\rho] - g_n(W)[\rho]| \le (e^{L\|\Delta\|_{\mathrm{F}}} - 1)|g_n(W)[\rho]|$. Summing over $n \in B$ and applying (146) proves (148); the identity $|a^2 - b^2| \le |a-b|(|a|+|b|)$ then gives (149). □

For the next lemma only, extend the reshuffling process to all integer epochs by taking independent uniform ordered partitions for every epoch index in $\mathbb{Z}$. For $z \ne \mathbf{0}$, define $Y_t(z) := T_{B_t} z/\sigma(z)$.

Lemma E.3 (Uniform concentration of frozen exponential averages)**.** *For every $\delta > 0$, there exists $\omega_\delta : (0,1) \to [0,1]$ with $\omega_\delta(\beta) \to 0$ as $\beta \uparrow 1$ such that, for every epoch start $\tau = rm$,*

$$\mathbb{P}\left(\max_{0\le s<m}\sup_{z\neq\mathbf{0}}\left|(1-\beta)\sum_{j=0}^{\infty}\beta^j\left(Y_{\tau+s-j}(z)-\mathbb{E}Y_0(z)\right)\right|>\delta\right)\le\omega_\delta(\beta).$$

The same statement holds with $Y_t(z)^2$ in place of $Y_t(z)$ and $\mathbb{E}[Y_0(z)^2] = 1$. For fixed N, b, one may take $\omega_\delta(\beta) \le C_\delta(1-\beta)$.

Proof. It is enough to work on the compact σ-unit sphere. Group the lags in the exponential average by reshuffling epoch. Contributions from distinct epochs are independent, every epoch block has mean zero, and Lemma E.1 uniformly bounds $|Y_t(z)|$. Hence the variance of the centered exponential average is at most $C(1-\beta)^2\sum_{q\ge0}\beta^{2qm} \le C'(1-\beta)$. Chebyshev's inequality gives the stated rate for each fixed z. The maps $z \mapsto T_B z$ are uniformly Lipschitz in the σ-norm, so a finite δ-net and a union bound make the estimate uniform over z and over $s = 0, \dots, m-1$. The argument for $Y_t(z)^2$ is identical because $Y_t(z)$ is uniformly bounded. □

For fixed W, define the two-sided frozen memories

$$\overline{M}_t^{\beta_1}(W) := (1-\beta_1)\sum_{j=0}^{\infty}\beta_1^j G_{B_{t-j}}(W), \qquad \overline{V}_t^{\beta_2}(W) := (1-\beta_2)\sum_{j=0}^{\infty}\beta_2^j G_{B_{t-j}}(W)^{\odot 2}.$$

Their expectations are $\mu_1(W)$ and $\mu_2(W)$, respectively.

Lemma E.4 (Freezing one epoch)**.** *Fix $0 < \beta_1 \le \beta_2 < 1$. Under the assumptions of Proposition 3.5, for every coordinate ρ, uniformly over $0 \le s < m$,*

$$\frac{|M_{\tau_r+s+1}[\rho] - \overline{M}^{\beta_1}_{\tau_r+s}(W_{\tau_r})[\rho]|}{\sigma_\rho(W_{\tau_r})} \xrightarrow[r\to\infty]{\mathbb{P}} 0, \tag{150}$$

$$\frac{|V_{\tau_r+s+1}[\rho] - \overline{V}^{\beta_2}_{\tau_r+s}(W_{\tau_r})[\rho]|}{\sigma_\rho(W_{\tau_r})^2} \xrightarrow[r\to\infty]{\mathbb{P}} 0. \tag{151}$$

The convergence is uniform over the finitely many coordinates.

Proof. Fix $t = \tau_r + s$. The first-moment difference consists of the missing negative-time tail and movement errors obtained by replacing W_{t-j} by W_{τ_r}. Lemma E.1 bounds the normalized missing tail by $C\beta_1^{t+1}$. For $0 \le j \le s$, Lemma E.2 and monotonicity of η_t give $\|W_{t-j} - W_{\tau_r}\|_{\mathrm{F}} \le C_{\beta_1,\beta_2} m\eta_{\tau_r} \to 0$. For $j > s$, writing $q = j - s$ gives $\|W_{\tau_r-q} - W_{\tau_r}\|_{\mathrm{F}} \le C_{\beta_1,\beta_2}\sum_{u=\tau_r-q}^{\tau_r-1}\eta_u$. Therefore (148) bounds the normalized past-lag contribution by

$$C(1-\beta_1)\sum_{q=1}^{\tau_r}\beta_1^{q+s}\left(\exp\left(C\sum_{u=\tau_r-q}^{\tau_r-1}\eta_u\right)-1\right),$$

which tends to zero by Assumption 2.6. This proves (150). The second-moment argument is identical using (149) and β_2, proving (151). □

Lemma E.5 (Continuity of the Adam ratio)**.** *There is a universal $C_{\mathrm{rat}} < \infty$ such that, whenever $\sigma > 0$, $|a_0| \le \sigma$, $|a - a_0| \le \delta\sigma$, and $|v - \sigma^2| \le \delta\sigma^2$ with $0 < \delta \le 1/2$,*

$$\left|\frac{a}{\sqrt{v}} - \frac{a_0}{\sigma}\right| \le C_{\mathrm{rat}}\delta.$$

Proof. Since $v \ge \sigma^2/2$,

$$\left|\frac{a}{\sqrt{v}} - \frac{a_0}{\sigma}\right| \le \frac{|a-a_0|}{\sqrt{v}} + |a_0|\left|\frac{1}{\sqrt{v}} - \frac{1}{\sigma}\right| \le \sqrt{2}\,\delta + \left|(1+e)^{-1/2} - 1\right|,$$

where $v = \sigma^2(1+e)$ and $|e| \le \delta$. The map $x \mapsto (1+x)^{-1/2}$ is Lipschitz on $[-1/2, 1/2]$. □

Proof of Proposition 3.5. Let $f_{N,b}(W) := \mu_1(W) \oslash \mu_2(W)^{\odot 1/2} = \Phi_{N,b}(W)$ by Lemma E.1. Because the Adam update uses the gradient convention,

$$\frac{W_{\tau_r+m} - W_{\tau_r}}{\gamma_r} = -\sum_{s=0}^{m-1} \frac{\eta_{\tau_r+s}}{\gamma_r} U_{\tau_r+s}, \qquad \sum_{s=0}^{m-1} \frac{\eta_{\tau_r+s}}{\gamma_r} = 1. \tag{152}$$

Fix $0 < \delta \le 1/4$. Lemma E.3 gives, with probability at least $1 - \omega_\delta(\beta_1) - \omega_\delta(\beta_2)$, relative δ-control of both frozen memories around $\mu_1(W_{\tau_r})$ and $\mu_2(W_{\tau_r})$, simultaneously over the epoch and all coordinates. Lemma E.4 transfers these bounds to the actual memories with 2δ in place of δ, up to an event whose probability vanishes as $r \to \infty$. Jensen's inequality gives $|\mu_1(W)[\rho]| \le \sqrt{\mu_2(W)[\rho]} = \sigma_\rho(W)$, so Lemma E.5 yields

$$\max_{0 \le s < m} \|U_{\tau_r+s} - \Phi_{N,b}(W_{\tau_r})\|_{\max} \le C\delta \tag{153}$$

with the same asymptotic probability. Combining (152) and (153) gives

$$\left\| \frac{W_{\tau_r+m} - W_{\tau_r}}{\gamma_r} + \Phi_{N,b}(W_{\tau_r}) \right\|_{\max} \le C\delta.$$

Given $\varepsilon > 0$, choose δ so that $C\delta < \varepsilon$. Taking $r \to \infty$ first yields a limsup probability bounded by $\omega_\delta(\beta_1) + \omega_\delta(\beta_2)$, and then taking $(\beta_1, \beta_2) \to (1, 1)$ through $0 < \beta_1 \le \beta_2 < 1$ proves (8). □

E.2 RR-RowSGDM versus AdamProxy population accuracy

We prove Proposition 3.6 by bounding the coordinates of the margin-normalized AdamProxy iterates and comparing the resulting classifier family with the near-row-margin family. All training-time limits are taken at fixed dimension before the limit $d \to \infty$.

A deterministic trajectory bound. Fix a finite linearly separable dataset $X = \{(x_n, y_n)\}_{n=1}^N$, with $x_n \in \mathbb{R}^d$ and $y_n \in [k]$. Write $p_n(W) = \mathbb{S}(Wx_n)$, $\ell_n(W) = -\log p_{n,y_n}(W)$, and $L(W) = N^{-1}\sum_{n=1}^N \ell_n(W)$. For $\mathcal{Q} := \{(n,c) : n \in [N],\ c \ne y_n\}$, set $A_{n,c} := (e_{y_n} - e_c)x_n^\top$, so that

$$\mathrm{mar}_X(W) = \min_{(n,c)\in\mathcal{Q}} \langle W, A_{n,c}\rangle, \qquad \gamma_{\max}(X) = \max_{\|V\|_{\max}\le 1} \mathrm{mar}_X(V) > 0.$$

Let $R_1(X) := \max_n \|x_n\|_1$ and $r^*_{\max}(X) = \gamma_{\max}(X)^{-1}$. For $\kappa \ge 0$, define

$$h_n(W) := (e_{y_n} - p_n(W))x_n^\top = \sum_{c \ne y_n} p_{n,c}(W) A_{n,c},$$

$$\widetilde{S}(W) := \sum_{n=1}^N h_n(W), \qquad \widetilde{Q}(W) := \sum_{n=1}^N h_n(W)^{\odot 2},$$

$$\Psi_\kappa(W) := \widetilde{S}(W) \oslash \left(\widetilde{S}(W)^{\odot 2} + \kappa \widetilde{Q}(W)\right)^{\odot 1/2}.$$

All powers and divisions are entrywise, with $0/0 := 0$.

Proposition E.6 (Positive margin growth and a trajectory-level coordinate bound)**.** *Let $\kappa \ge 0$, let $W_0 \in \mathbb{R}^{k\times d}$ be finite, and consider $W_{t+1} = W_t + a_t\Psi_\kappa(W_t)$, where $a_t > 0$, $a_t \to 0$, and $\sum_{t=0}^\infty a_t = \infty$. Define*

$$T_t := \sum_{s=0}^{t-1} a_s, \qquad K_\kappa(X) := 2(1+\sqrt{\kappa})R_1(X), \qquad \rho_\kappa(X) := \frac{\gamma_{\max}(X)^2}{K_\kappa(X)} > 0.$$

Then $L(W_t) \to 0$, $\mathrm{mar}_X(W_t) \to \infty$, and

$$\liminf_{t\to\infty} \frac{\mathrm{mar}_X(W_t)}{T_t} \ge \rho_\kappa(X), \tag{154}$$

$$\liminf_{t\to\infty} \frac{\mathrm{mar}_X(W_t)}{\|W_t\|_{\max}} \ge \rho_\kappa(X), \tag{155}$$

$$\limsup_{t\to\infty} \left\| \frac{W_t}{\mathrm{mar}_X(W_t)} \right\|_{\max} \le K_\kappa(X) r^*_{\max}(X)^2. \tag{156}$$

The margin-normalized matrices in (156) *are well defined and have minimum training margin one for all sufficiently large t.*

Proof. For finite W, define

$$\mathcal{G}(W) := \frac{1}{N}\sum_{n=1}^{N}(1 - p_{n,y_n}(W)), \qquad R(W) := \sum_{(n,c)\in\mathcal{Q}} p_{n,c}(W) = N\mathcal{G}(W) > 0,$$

and let $\theta_{n,c}(W) := p_{n,c}(W)/R(W)$. Then $\theta(W)$ belongs to the probability simplex $\Delta(\mathcal{Q})$. For any $\theta \in \Delta(\mathcal{Q})$, write

$$Z_n(\theta) := \sum_{c\neq y_n} \theta_{n,c} A_{n,c}, \qquad \mathsf{S}(\theta) := \sum_{n=1}^{N} Z_n(\theta),$$
$$\mathsf{D}_\kappa(\theta) := \left(\mathsf{S}(\theta)^{\odot 2} + \kappa \sum_{n=1}^{N} Z_n(\theta)^{\odot 2}\right)^{\odot 1/2}.$$

Since $h_n(W) = R(W)Z_n(\theta(W))$, the common factor $R(W)$ cancels in the proxy ratio, giving

$$\Psi_\kappa(W) = \mathsf{S}(\theta(W)) \oslash \mathsf{D}_\kappa(\theta(W)), \qquad \|\Psi_\kappa(W)\|_{\max} \le 1. \tag{157}$$

The coordinate bound follows from $\mathsf{D}_\kappa(\theta)[\rho] \ge |\mathsf{S}(\theta)[\rho]|$, including zero denominators under the stated convention. The entrywise inequalities $|\mathsf{S}(\theta)| \le \sum_n |Z_n(\theta)|$ and $(\sum_n Z_n(\theta)^{\odot 2})^{\odot 1/2} \le \sum_n |Z_n(\theta)|$ imply

$$\|\mathsf{D}_\kappa(\theta)\|_{\mathrm{sum}} \le (1+\sqrt{\kappa})\sum_n \|Z_n(\theta)\|_{\mathrm{sum}} \le 2(1+\sqrt{\kappa})R_1(X) = K_\kappa(X).$$

Here the last inequality uses $\|A_{n,c}\|_{\mathrm{sum}} = 2\|x_n\|_1$ and $\sum_{(n,c)\in\mathcal{Q}} \theta_{n,c} = 1$. Let V_* attain $\gamma_{\max}(X)$ on the compact max-norm unit ball. Max–sum duality and the margin constraints for V_* give

$$\|\mathsf{S}(\theta)\|_{\mathrm{sum}} \ge \langle V_*, \mathsf{S}(\theta)\rangle = \sum_{(n,c)\in\mathcal{Q}} \theta_{n,c}\langle V_*, A_{n,c}\rangle \ge \gamma_{\max}(X).$$

If $\mathsf{D}_\kappa(\theta)[\rho] = 0$, then $\mathsf{S}(\theta)[\rho] = 0$. Weighted Cauchy–Schwarz on the remaining coordinates therefore yields

$$\mathcal{J}_\kappa(\theta) := \sum_{\rho:\mathsf{D}_\kappa(\theta)[\rho]>0} \frac{\mathsf{S}(\theta)[\rho]^2}{\mathsf{D}_\kappa(\theta)[\rho]} \ge \frac{\|\mathsf{S}(\theta)\|_{\mathrm{sum}}^2}{\|\mathsf{D}_\kappa(\theta)\|_{\mathrm{sum}}} \ge \rho_\kappa(X).$$

Since $\nabla L(W) = -\mathcal{G}(W)\mathsf{S}(\theta(W))$, it follows that

$$-\langle \nabla L(W), \Psi_\kappa(W)\rangle = \mathcal{G}(W)\mathcal{J}_\kappa(\theta(W)) \ge \rho_\kappa(X)\mathcal{G}(W). \tag{158}$$

To control the discrete-time remainder, fix $\|\Delta\|_{\max} \le 1$. Then $\|\Delta x_n\|_\infty \le R_1(X)$, and the softmax Hessian satisfies

$$D^2\ell_n(V)[\Delta,\Delta] = \mathrm{Var}_{C\sim p_n(V)}\big((\Delta x_n)_C\big) \le 4R_1(X)^2\big(1 - p_{n,y_n}(V)\big).$$

Indeed, the variance is at most the second moment about $(\Delta x_n)_{y_n}$, and every logit difference has magnitude at most $2R_1(X)$. The softmax formula also gives $\mathcal{G}(W + sa\Delta) \le e^{2R_1(X)a}\mathcal{G}(W)$ for $a \ge 0$ and $s \in [0,1]$. Because $a_t \to 0$, the quantity $a_{\max} := \sup_t a_t$ is finite. Taylor's formula, (157), and (158) consequently imply

$$L(W_{t+1}) \le L(W_t) - a_t\big(\rho_\kappa(X) - C_X a_t\big)\mathcal{G}(W_t), \qquad C_X := 2R_1(X)^2 e^{2R_1(X)a_{\max}}. \tag{159}$$

For all sufficiently large t, the coefficient in parentheses is at least $\rho_\kappa(X)/2$, so $L(W_t)$ is eventually nonincreasing and has a nonnegative limit. If this limit were $L_\infty > 0$, then at each sufficiently large t some sample would satisfy $\ell_n(W_t) \ge L(W_t) \ge L_\infty$, and hence $\mathcal{G}(W_t) \ge (1 - e^{-L_\infty})/N > 0$. Summing (159) would contradict $L \ge 0$ and $\sum_t a_t = \infty$. Thus $L(W_t) \to 0$. Since $0 \le \ell_n(W_t) \le NL(W_t)$, every sample loss tends to zero, and

$$\frac{\mathcal{G}(W_t)}{L(W_t)} = \frac{\sum_n (1 - e^{-\ell_n(W_t)})}{\sum_n \ell_n(W_t)} \longrightarrow 1.$$

Moreover,

$$\ell_n(W) = \log\left(1 + \sum_{c \neq y_n} e^{-\langle W, A_{n,c}\rangle}\right)$$

shows that every training margin tends to infinity. For all sufficiently large t, dividing (159) by $L(W_t)$ and applying $-\log(1-z) \geq z$ gives

$$-\log L(W_{t+1}) + \log L(W_t) \geq a_t\big(\rho_\kappa(X) - C_X a_t\big)\frac{\mathcal{G}(W_t)}{L(W_t)}.$$

The factor multiplying a_t converges to $\rho_\kappa(X)$, so summation and weighted Cesàro averaging yield

$$\liminf_{t\to\infty} \frac{-\log L(W_t)}{T_t} \geq \rho_\kappa(X).$$

Write $m_t := \operatorname{mar}_X(W_t)$. Once $m_t \geq 0$, a sample attaining the minimum margin gives

$$L(W_t) \geq \frac{1}{N}\log(1 + e^{-m_t}) \geq \frac{e^{-m_t}}{2N}, \qquad m_t \geq -\log L(W_t) - \log(2N).$$

This proves (154). Finally, (157) gives $\|W_t\|_{\max} \leq \|W_0\|_{\max} + T_t$. Combining this bound with (154) proves (155) and (156). □

For $2 \leq b \leq N$, the explicit AdamProxy formula gives

$$-\Phi_{N,b}(W) = c^{\mathrm{AP}}_{N,b}\Psi_{\kappa_{N,b}}(W), \qquad c^{\mathrm{AP}}_{N,b} := \sqrt{\frac{b(N-1)}{N(b-1)}}, \qquad \kappa_{N,b} := \frac{N-b}{b-1}.$$

Thus (9) has the recursion in Proposition E.6 with $a_t = c^{\mathrm{AP}}_{N,b}\gamma_t$ and $\gamma_t = \sum_{s=0}^{m-1}\eta_{mt+s}$. At each fixed dimension, Assumption 2.5 implies $a_t \to 0$ and $\sum_t a_t = \infty$. The full-batch case $b = N$ corresponds to $\kappa_{N,b} = 0$, $c^{\mathrm{AP}}_{N,b} = 1$, and $\Psi_0 = \operatorname{sign}(\widetilde{S})$.

Gaussian-cloud normalization and coordinate caps. We use the normalized convention (38), under which $\mu_i, z_{i,s} \sim \mathcal{N}(\mathbf{0}, I_d/d)$ and $X^{\mathrm{te}} \mid (Y = i) \sim \mathcal{N}(\mu_i, r_{\mathrm{te}}^2 I_d/d)$. This convention also preserves the relevant trajectory estimates. Indeed, if $X^{\mathrm{raw}} = \sqrt{d}\,X$ and $V = \sqrt{d}\,W$, then every normalized-data sample gradient at V equals $d^{-1/2}$ times the corresponding raw-data gradient at W, so $\Psi_{\kappa,X}(V) = \Psi_{\kappa,X^{\mathrm{raw}}}(W)$. Consequently, $V_t = \sqrt{d}\,W_t$ follows the normalized-data recursion with step sizes $\sqrt{d}\,a_t$ and identical predictions. For RR-RowSGDM, scaling its momentum by $d^{-1/2}$ gives the same row-normalized update direction, so its transformed parameter trajectory uses step sizes $\sqrt{d}\,\eta_t$. Multiplication by this fixed positive constant preserves Assumptions 2.5 and 2.6 at each fixed d. We suppress the rescaling in the notation below.

Under the assumptions of Proposition 3.6, $b_d \geq N(d)/m_0 \to \infty$, so $b_d \geq 2$ for all sufficiently large d. Moreover,

$$0 \leq \kappa_{N,b_d} = \frac{b_d(m_d - 1)}{b_d - 1} \leq 2(m_0 - 1) =: \overline{\kappa}. \tag{160}$$

Set

$$R_* := 2(1 + 2r_0), \qquad C_{\mathrm{AP}} := \pi R_*(1 + \sqrt{\overline{\kappa}}), \qquad a_{\mathrm{AP}} := \max\{3, 2C_{\mathrm{AP}} + 1\},$$

and define the capped feasible family

$$\mathcal{C}_d := \left\{V \in \mathcal{F}_X : \|V\|_{\max} \leq \frac{C_{\mathrm{AP}} + 1/2}{\sqrt{d}}\right\}.$$

For the functions $A_{\max}$ and $B_{\max}$ in (39), and for $a \geq 3$, let $\tau_a \in (0,1)$ be the unique solution of

$$A_{\max}(\tau_a) = \frac{\sqrt{2}}{a}.$$

Such a solution exists because $A_{\max}(0) = 0$, $A_{\max}$ is strictly increasing, and $A_{\max}(1) > 1/\sqrt{\pi} > \sqrt{2}/a$. Define

$$C_{\mathrm{clip}}(a) := \frac{\sqrt{2}}{a\sqrt{B_{\max}(\tau_a)}}, \qquad \delta_{\mathrm{clip}}(a) := \sqrt{2 - 2C_{\mathrm{clip}}(a)}, \qquad \delta_{\mathrm{AP}} := \delta_{\mathrm{clip}}(a_{\mathrm{AP}}).$$

Strict Cauchy–Schwarz gives $A_{\max}(\tau)^2 < B_{\max}(\tau)$ for every finite $\tau > 0$, so $0 < C_{\mathrm{clip}}(a) < 1$ and $\delta_{\mathrm{AP}} > 0$.

Lemma E.7 (Uniform clipped-Gaussian angular barrier)**.** *Let $u_{d,\xi} \in \mathbb{S}^{d-1}$ be indexed by a nonempty finite set $\mathcal{I}_d$, and suppose the empirical functions in* (48) *satisfy*

$$\sup_{\xi\in\mathcal{I}_d}\sup_{\tau\in[0,1]} \left(\left|H_{u_{d,\xi}}(\tau) - A_{\max}(\tau)\right| + \left|J_{u_{d,\xi}}(\tau) - B_{\max}(\tau)\right|\right) \longrightarrow 0.$$

Fix $a \geq 3$, let $a_d \to a$ and $\upsilon_d \to \sqrt{2}$ be positive deterministic sequences, and define

$$\mathcal{V}_{d,\xi} := \left\{ v \in \mathbb{R}^d : \|v\|_\infty \leq \frac{a_d}{\sqrt{d}}, \quad v^\top u_{d,\xi} \geq \upsilon_d \right\}.$$

For all sufficiently large d, these sets are nonempty and

$$\inf_{\xi\in\mathcal{I}_d}\inf_{v\in\mathcal{V}_{d,\xi}} \left\| \frac{v}{\|v\|_2} - u_{d,\xi} \right\|_2 \geq \delta_{\mathrm{clip}}(a) - o_d(1). \tag{161}$$

The error depends only on the uniform empirical errors and the deviations of a_d and υ_d from their limits.

Proof. Since $\upsilon_d/a_d \to \sqrt{2}/a < A_{\max}(1)$, uniform convergence gives $0 < \upsilon_d/a_d < H_{u_{d,\xi}}(1)$ for every ξ and all sufficiently large d. The feasibility condition in Proposition C.8 therefore holds, and that proposition gives

$$\mathfrak{m}_{d,\xi} := \min_{v\in\mathcal{V}_{d,\xi}} \|v\|_2 = a_d\sqrt{J_{u_{d,\xi}}(\widehat{\tau}_{d,\xi})}, \qquad H_{u_{d,\xi}}(\widehat{\tau}_{d,\xi}) = \frac{\upsilon_d}{a_d},$$

with $\widehat{\tau}_{d,\xi} \in (0,1)$. Uniform convergence and strict monotonicity of $A_{\max}$ imply

$$\sup_{\xi\in\mathcal{I}_d} |\widehat{\tau}_{d,\xi} - \tau_a| \longrightarrow 0, \qquad \sup_{\xi\in\mathcal{I}_d} \left|\mathfrak{m}_{d,\xi} - a\sqrt{B_{\max}(\tau_a)}\right| \longrightarrow 0.$$

For every $v \in \mathcal{V}_{d,\xi}$, the vector $y = (\upsilon_d/(v^\top u_{d,\xi}))v$ also belongs to $\mathcal{V}_{d,\xi}$. Hence

$$\frac{v^\top u_{d,\xi}}{\|v\|_2} \leq \frac{\upsilon_d}{\mathfrak{m}_{d,\xi}} = C_{\mathrm{clip}}(a) + o_d(1),$$

where the error is independent of v and uniform over ξ. The identity $\|v/\|v\|_2 - u\|_2^2 = 2 - 2v^\top u/\|v\|_2$ proves (161). □

Proposition E.8 (Capped AdamProxy trajectories and boundary separation)**.** *For every sufficiently large d, on the event $\mathcal{E}_d^{\mathrm{near}}$ from Lemma D.2, the AdamProxy trajectory satisfies*

$$\frac{W_t^{\mathrm{AP}}}{\mathrm{mar}_X(W_t^{\mathrm{AP}})} \in \mathcal{C}_d \qquad \textit{for all sufficiently large } t, \tag{162}$$

and

$$\inf_{V\in\mathcal{C}_d}\min_{i\neq j} \|u_{ij}^V - u_{ij}^\mu\|_2 \geq \delta_{\mathrm{AP}} - o_d(1). \tag{163}$$

The error is deterministic on $\mathcal{E}_d^{\mathrm{near}}$ and uniform over $V \in \mathcal{C}_d$. The time threshold in (162) *may depend on the dimension, dataset, initialization, and learning-rate schedule.*

Proof. The event guarantees separability, and Proposition C.4(H5) and (128) give

$$r^*_{\max}(X) \leq \frac{\sqrt{2\pi} + o_d(1)}{2\sqrt{d}}, \qquad \max_n \|x_n\|_2 \leq R_*, \qquad R_1(X) \leq \sqrt{d}\, R_*.$$

Applying Proposition E.6 and (160), we obtain

$$\limsup_{t\to\infty}\sqrt{d}\left\|\frac{W_t^{\mathrm{AP}}}{\mathrm{mar}_X(W_t^{\mathrm{AP}})}\right\|_{\max} \le 2(1+\sqrt{\kappa})dR_*r^*_{\max}(X)^2 \le C_{\mathrm{AP}}+o_d(1).$$

For all sufficiently large d, the right-hand side is less than $C_{\mathrm{AP}}+1/4$. The definition of the time limsup and the unit margin of the normalized iterates then give (162). In particular, $\mathcal{C}_d$ is nonempty.

Fix $V\in\mathcal{C}_d$ and an ordered pair $i\neq j$, and write $v:=v_i-v_j$ and $u:=u_{ij}^m$. The coordinate cap gives

$$\|v\|_\infty \le \frac{2C_{\mathrm{AP}}+1}{\sqrt{d}} \le \frac{a_{\mathrm{AP}}}{\sqrt{d}}.$$

Averaging feasibility over the samples from classes i and j gives $v^\top m_i\ge 1$ and $v^\top m_j\le -1$, and hence

$$v^\top u \ge \frac{2}{\|m_i-m_j\|_2} \ge \sqrt{2}-o_d(1).$$

The last error is deterministic and uniform over all pairs and all $V\in\mathcal{C}_d$, by Proposition C.4(H2). Choose a positive deterministic sequence $\upsilon_d\to\sqrt{2}$ below these projection lower bounds. Proposition C.4(H6) supplies the empirical estimates in Lemma E.7, which applies with $a_d=a_{\mathrm{AP}}$ and $\mathcal{I}_d=\{(i,j):i\neq j\}$. Thus $\|u_{ij}^V-u_{ij}^m\|_2\ge\delta_{\mathrm{AP}}-o_d(1)$ uniformly over all pairs and all $V\in\mathcal{C}_d$. Finally, Proposition C.4(H4) and the reverse triangle inequality prove (163). □

High-momentum row geometry. Let C_{row} be the constant in Proposition D.3, and let K_{row} be the row constant in Proposition D.4. Choose $h_{\mathrm{AP}}:=\delta_{\mathrm{AP}}^2/(32\sqrt{2})$, so that

$$\mathsf{a}(h_{\mathrm{AP}})>\mathsf{b}(\delta_{\mathrm{AP}}/2), \tag{164}$$

where a and b are defined in (98). Set

$$b_{\mathrm{AP}}:=\min\left\{1,\left(\frac{h_{\mathrm{AP}}}{C_{\mathrm{row}}}\right)^4\right\},\qquad \vartheta_{\mathrm{AP}}:=\frac{b_{\mathrm{AP}}/2}{1+b_{\mathrm{AP}}/2},\qquad A_0:=m_0(m_0^2-1),$$

and

$$\beta_{\mathrm{AP}}:=\begin{cases}\dfrac{1}{2}, & A_0=0,\\ \max\left\{\dfrac{1}{2},1-\dfrac{\vartheta_{\mathrm{AP}}}{2A_0K_{\mathrm{row}}}\right\}, & A_0>0.\end{cases} \tag{165}$$

These constants depend only on (α,r_0,m_0), and $\beta_{\mathrm{AP}}\in(0,1)$.

Lemma E.9 (High-momentum Row near-set control)**.** *For every fixed $\beta\in[\beta_{\mathrm{AP}},1)$, on $\mathcal{E}_d^{\mathrm{near}}$ and for every sufficiently large d, the RR-RowSGDM effective margin is positive and*

$$\sup_{U\in\mathcal{U}^X_{\mathrm{row},\beta,b_d}}\ \max_{i\neq j}\|u_{ij}^U-u_{ij}^\mu\|_2\le h_{\mathrm{AP}}.$$

Proof. The relative margin deficit is

$$\Delta_{\mathrm{row},d}:=\frac{\varepsilon_{\mathrm{row},\beta,b_d}(X)}{\gamma_{\mathrm{row}}(X)}=2(1-\beta)m_d(m_d^2-1)\frac{R_{\mathrm{row}}(X)}{\gamma_{\mathrm{row}}(X)}.$$

Proposition D.4 and (165) give $\Delta_{\mathrm{row},d}\le 2(1-\beta)A_0K_{\mathrm{row}}\le\vartheta_{\mathrm{AP}}<1$, including $\Delta_{\mathrm{row},d}=0$ when $m_d=1$. Thus the effective margin is positive, and

$$\frac{\Delta_{\mathrm{row},d}}{1-\Delta_{\mathrm{row},d}}\le\frac{b_{\mathrm{AP}}}{2}.$$

Lemma D.1 maps every $U\in\mathcal{U}^X_{\mathrm{row},\beta,b_d}$ to a unit-margin feasible representative with the same pairwise row directions and relative row-norm slack at most $b_{\mathrm{AP}}/2$. For all sufficiently large d, $\omega_d\le b_{\mathrm{AP}}/2$, so Proposition D.3 gives the uniform bound $C_{\mathrm{row}}b_{\mathrm{AP}}^{1/4}\le h_{\mathrm{AP}}$. □

Proof of Proposition 3.6. Set $\beta_1 := \beta_{\mathrm{AP}}$, fix $\beta \in [\beta_1, 1)$, and let $\mathcal{R}_d$ be the Gaussian radial event from Lemma C.18. Define $\mathcal{H}_d := \mathcal{E}_d^{\mathrm{near}} \cap \mathcal{R}_d$. By (127) and (95), for every sufficiently large d,

$$\mathbb{P}(\mathcal{H}_d) \geq 1 - e^{-\sqrt{d}}.$$

Work on $\mathcal{H}_d$, and take

$$\mathcal{R}_d^{\mathrm{clf}} := \mathcal{U}^X_{\mathrm{row},\beta,b_d}, \qquad \mathcal{C}_d^{\mathrm{clf}} := \mathcal{C}_d.$$

Lemma E.9 and Proposition E.8 give, for all sufficiently large d,

$$\sup_{U \in \mathcal{R}_d^{\mathrm{clf}}} \max_{i \neq j} \|u_{ij}^U - u_{ij}^\mu\|_2 \leq h_{\mathrm{AP}}, \qquad \inf_{V \in \mathcal{C}_d} D_{\text{avg-bdry}}(V) \geq \frac{\delta_{\mathrm{AP}}}{2}.$$

The radial deviations tend to zero on $\mathcal{R}_d$ under a deterministic envelope, and the limiting amplitude condition is (164). Choose fixed constants $0 < b_* < a_*$ with $\mathsf{b}(\delta_{\mathrm{AP}}/2) < b_* < a_* < \mathsf{a}(h_{\mathrm{AP}})$. Proposition C.19 gives, for $\sigma_d = r_{\mathrm{te}}/\sqrt{d}$ and all sufficiently large d,

$$\sup_{U \in \mathcal{U}^X_{\mathrm{row},\beta,b_d}} \mathcal{E}_{\pi,d}(U) \leq (k-1)\overline{\Phi}(a_*/\sigma_d),$$

$$\inf_{V \in \mathcal{C}_d} \mathcal{E}_{\pi,d}(V) \geq 2\pi_{\min,d}\overline{\Phi}(b_*/\sigma_d).$$

The Gaussian upper-tail inequality $\overline{\Phi}(x)/\overline{\Phi}(y) \leq e^{-(x^2-y^2)/2}$ for $x \geq y \geq 0$ therefore yields

$$\frac{\sup_{U \in \mathcal{U}^X_{\mathrm{row},\beta,b_d}} \mathcal{E}_{\pi,d}(U)}{\inf_{V \in \mathcal{C}_d} \mathcal{E}_{\pi,d}(V)} \leq \frac{k-1}{2\pi_{\min,d}} \exp\left(-\frac{(a_*^2 - b_*^2)d}{2r_{\mathrm{te},0}^2}\right) \longrightarrow 0,$$

since $k = O(d)$, $\log(1/\pi_{\min,d}) = o(d)$, and $r_{\mathrm{te},0} := \sup_d r_{\mathrm{te}}(d) < \infty$. Consequently, for every sufficiently large d on $\mathcal{H}_d$,

$$\sup_{U \in \mathcal{U}^X_{\mathrm{row},\beta,b_d}} \mathcal{E}_{\pi,d}(U) < \inf_{V \in \mathcal{C}_d} \mathcal{E}_{\pi,d}(V). \tag{166}$$

The positive Row effective margin and Corollary B.12 give

$$d_{\cos}\big(W_t^{\text{RR-RowSGDM}}, \mathcal{U}^X_{\mathrm{row},\beta,b_d}\big) \longrightarrow 0.$$

Lemma C.21 hence implies

$$\limsup_{t\to\infty} \mathcal{E}_{\pi,d}\big(W_t^{\text{RR-RowSGDM}}\big) \leq \sup_{U \in \mathcal{U}^X_{\mathrm{row},\beta,b_d}} \mathcal{E}_{\pi,d}(U).$$

For AdamProxy, (162) and positive scale invariance of population error give

$$\liminf_{t\to\infty} \mathcal{E}_{\pi,d}(W_t^{\mathrm{AP}}) \geq \inf_{V \in \mathcal{C}_d} \mathcal{E}_{\pi,d}(V).$$

Combining these two bounds with (166) yields

$$\limsup_{t\to\infty} \mathcal{E}_{\pi,d}\big(W_t^{\text{RR-RowSGDM}}\big) < \liminf_{t\to\infty} \mathcal{E}_{\pi,d}(W_t^{\mathrm{AP}}).$$

Since $\mathrm{Acc}_{\pi,d} = 1 - \mathcal{E}_{\pi,d}$, this is the claimed accuracy ordering. The Row momentum estimates hold for every reshuffling realization, and the AdamProxy estimates are deterministic conditional on the training data, so no additional exceptional event is required. □

E.3 AdamProxy margin characterization

AdamProxy is not steepest descent in a fixed norm, so we characterize its limit through a data-dependent quadratic margin problem. Let $\mathcal{Q} := \{(n,c) : n \in [N], c \neq y_n\}$, let $\Delta(\mathcal{Q})$ be the probability simplex, and for $\theta \in \Delta(\mathcal{Q})$ define

$$Z_n(\theta) := \sum_{c \neq y_n} \theta_{n,c}(e_{y_n} - e_c)x_n^\top,$$

$$M_{N,b}(\theta) := \mathrm{diag}\left(\mathrm{vec}\left[\left(\frac{N}{b(N-1)}\left((b-1)\left(\sum_{n=1}^N Z_n(\theta)\right)^{\odot 2} + (N-b)\sum_{n=1}^N Z_n(\theta)^{\odot 2}\right)\right)^{\odot 1/2}\right]\right).$$

Extending the parametric margin formulation of Baek et al. (2026), let

$$\mathcal{P}_{N,b}(\theta) := \arg\min_{W} \frac{1}{2} \operatorname{vec}(W)^\top M_{N,b}(\theta) \operatorname{vec}(W) \quad \text{s.t.} \quad (w_{y_n} - w_c)^\top x_n \geq 1, \ \ (n,c) \in \mathcal{Q}. \tag{167}$$

Proposition E.10 (AdamProxy margin characterization)**.** *Under Assumption 2.5, let W_t follow (9) with $1 \leq b \leq N$. Assume linear separability, $x_n[a] \neq 0$ for all n, a, and $\Phi_{N,b}(W_t)/\|\Phi_{N,b}(W_t)\|_{\mathrm{F}} \to \overline{\Phi}$. Then $W_t/\|W_t\|_{\mathrm{F}} \to \overline{W} := -\overline{\Phi}$ and $\gamma_{\mathrm{AP}} := \max_X(\overline{W}) > 0$. For $p^\star := \overline{W}/\gamma_{\mathrm{AP}}$, there exist $\theta^\star \in \Delta(\mathcal{Q})$ and a nonzero dual optimizer $\lambda^\star \geq 0$ of the corresponding problem such that*

$$p^\star \in \mathcal{P}_{N,b}(\theta^\star), \qquad \theta^\star = \lambda^\star/\|\lambda^\star\|_1.$$

Thus, the AdamProxy margin direction is characterized by $\mathcal{P}_{N,b}(\theta^\star)$ at a self-consistent parameter $\theta^\star$, which coincides with the normalized dual weights of the same variational problem.

Proof setup. The remainder of this subsection proves Proposition E.10. We first rewrite the AdamProxy dynamics in terms of descent signals and normalized wrong-class coefficients, then characterize the trajectory-selected limiting coefficients, and finally verify the primal–dual optimality conditions of the quadratic margin problem.

Define the descent signals

$$h_n(W) := -g_n(W) = (e_{y_n} - p_n(W))x_n^\top,$$

and write $A_{n,c} := (e_{y_n} - e_c)x_n^\top$ for $(n,c) \in \mathcal{Q}$. Then $h_n(W) = \sum_{c \neq y_n} p_{n,c}(W) A_{n,c}$. Set

$$\widetilde{S}(W) := \sum_{n=1}^N h_n(W) = -S(W), \qquad \widetilde{Q}(W) := \sum_{n=1}^N h_n(W)^{\odot 2} = Q(W).$$

For $\theta \in \Delta(\mathcal{Q})$, write

$$\mathsf{S}(\theta) := \sum_{n=1}^N Z_n(\theta), \qquad \mathsf{Q}(\theta) := \sum_{n=1}^N Z_n(\theta)^{\odot 2},$$

$$\mathsf{D}_{N,b}(\theta) := \left(\frac{N}{b(N-1)} \left((b-1)\mathsf{S}(\theta)^{\odot 2} + (N-b)\mathsf{Q}(\theta) \right) \right)^{\odot 1/2}.$$

Thus $M_{N,b}(\theta) = \operatorname{diag}(\operatorname{vec} \mathsf{D}_{N,b}(\theta))$. Define

$$\Psi_{N,b}(W) := \widetilde{S}(W) \oslash \left(\frac{N}{b(N-1)} \left((b-1)\widetilde{S}(W)^{\odot 2} + (N-b)\widetilde{Q}(W) \right) \right)^{\odot 1/2},$$

with $0/0 := 0$ entrywise. The explicit proxy formula in the main text gives $\Psi_{N,b}(W) = -\Phi_{N,b}(W)$, so the AdamProxy dynamics can be written as $W_{t+1}^{\mathrm{AP}} = W_t^{\mathrm{AP}} + \gamma_t \Psi_{N,b}(W_t^{\mathrm{AP}})$.

Lemma E.11 (Coefficient representation and trajectory-selected limit)**.** *For finite W, define*

$$R(W) := \sum_{(n,c) \in \mathcal{Q}} p_{n,c}(W) > 0, \qquad \theta_{n,c}(W) := \frac{p_{n,c}(W)}{R(W)}.$$

Then

$$h_n(W) = R(W) Z_n(\theta(W)), \qquad \widetilde{S}(W) = R(W)\mathsf{S}(\theta(W)), \qquad \widetilde{Q}(W) = R(W)^2 \mathsf{Q}(\theta(W)),$$
$$\Psi_{N,b}(W) = \mathsf{S}(\theta(W)) \oslash \mathsf{D}_{N,b}(\theta(W)), \qquad \|\Psi_{N,b}(W)\|_{\max} \leq 1. \tag{168}$$

Under the assumptions of Proposition E.10,

$$\frac{W_t^{\mathrm{AP}}}{\|W_t^{\mathrm{AP}}\|_{\mathrm{F}}} \longrightarrow \overline{W} := -\overline{\Phi}, \tag{169}$$

and there exist a subsequence $t_j \to \infty$, $\theta^\star \in \Delta(\mathcal{Q})$, and $\nu \in (0, \infty)$ such that

$$\theta(W_{t_j}^{\mathrm{AP}}) \to \theta^\star, \qquad \|\Psi_{N,b}(W_{t_j}^{\mathrm{AP}})\|_{\mathrm{F}} \to \nu, \qquad \Psi_{N,b}(W_{t_j}^{\mathrm{AP}}) \to \nu\overline{W}, \tag{170}$$

with $\mathsf{S}(\theta^\star) \neq \mathbf{0}$ and $\nu\|\overline{W}\|_{\max} \leq 1$.

Proof. The factorization follows directly from $h_n(W) = \sum_{c \neq y_n} p_{n,c}(W) A_{n,c}$. For every coordinate ρ, Cauchy–Schwarz gives $\mathsf{S}(\theta)[\rho]^2 \le N\mathsf{Q}(\theta)[\rho]$, and therefore

$$\mathsf{D}_{N,b}(\theta)[\rho]^2 - \mathsf{S}(\theta)[\rho]^2 = \frac{N-b}{b(N-1)} \left(N\mathsf{Q}(\theta)[\rho] - \mathsf{S}(\theta)[\rho]^2 \right) \ge 0.$$

Hence $\mathsf{D}_{N,b}(\theta)[\rho] \ge |\mathsf{S}(\theta)[\rho]|$, and if $\mathsf{D}_{N,b}(\theta)[\rho] = 0$, then $\mathsf{S}(\theta)[\rho] = 0$. With the convention $0/0 := 0$, this proves (168) and the coordinate bound. By linear separability, choose V and $\sigma_{\mathrm{sep}} > 0$ with $\langle V, A_q \rangle \ge \sigma_{\mathrm{sep}}$ for every q. Since $\mathsf{D}_{N,b}$ is continuous on the compact simplex, $D_{\max} := \max_\theta \|\mathsf{D}_{N,b}(\theta)\|_{\max} < \infty$, and

$$\sigma_{\mathrm{sep}} \le \langle V, \mathsf{S}(\theta(W)) \rangle \le \|V\|_{\mathrm{F}} D_{\max} \|\Psi_{N,b}(W)\|_{\mathrm{F}}.$$

Thus $\|\Psi_{N,b}(W_t)\|_{\mathrm{F}}$ is bounded above and below by positive constants. Since $\Psi_{N,b}(W) = -\Phi_{N,b}(W)$, the normalized descent-proxy directions satisfy $\Psi_{N,b}(W_t)/\|\Psi_{N,b}(W_t)\|_{\mathrm{F}} \to -\overline{\Phi}$. Assumption 2.5 implies that the epochwise schedule γ_t is positive, tends to zero, and satisfies $\sum_t \gamma_t = \infty$. With $T_t := \sum_{s=0}^{t-1} \gamma_s \|\Psi_{N,b}(W_s)\|_{\mathrm{F}}$, one has $T_t \to \infty$ and

$$\frac{W_t}{T_t} = \frac{W_0}{T_t} + \frac{1}{T_t} \sum_{s=0}^{t-1} \gamma_s \|\Psi_{N,b}(W_s)\|_{\mathrm{F}} \frac{\Psi_{N,b}(W_s)}{\|\Psi_{N,b}(W_s)\|_{\mathrm{F}}} \longrightarrow -\overline{\Phi}.$$

Hence $\|W_t\|_{\mathrm{F}}/T_t \to 1$ and $W_t/\|W_t\|_{\mathrm{F}} \to -\overline{\Phi}$, proving (169). Compactness of $\Delta(\mathcal{Q})$ and the two-sided bound on $\|\Psi_{N,b}(W_t)\|_{\mathrm{F}}$ give a subsequence satisfying (170). The separator inequality passes to the limit and gives $\mathsf{S}(\theta^\star) \neq \mathbf{0}$, while $\|\Psi_{N,b}(W_{t_j})\|_{\max} \le 1$ yields $\nu \|\overline{W}\|_{\max} \le 1$. □

Proposition E.12 (Selected stationarity and positive margin)**.** *For the limit in Lemma E.11, there exists $\alpha = \nu^{-1} > 0$ such that*

$$M_{N,b}(\theta^\star) \operatorname{vec}(\overline{W}) = \alpha \operatorname{vec}(\mathsf{S}(\theta^\star)) = \alpha \sum_{q \in \mathcal{Q}} \theta_q^\star \operatorname{vec}(A_q). \tag{171}$$

Moreover,

$$\gamma_{\mathrm{AP}} := \min_{q \in \mathcal{Q}} \langle \overline{W}, A_q \rangle > 0, \qquad \operatorname{supp}(\theta^\star) \subseteq \arg\min_{q \in \mathcal{Q}} \langle \overline{W}, A_q \rangle. \tag{172}$$

Proof. Let $\mathsf{S}^\star := \mathsf{S}(\theta^\star)$ and $\mathsf{D}^\star := \mathsf{D}_{N,b}(\theta^\star)$. If $\mathsf{D}^\star[\rho] > 0$, continuity and (170) give $\mathsf{S}^\star[\rho]/\mathsf{D}^\star[\rho] = \nu \overline{W}[\rho]$. If $\mathsf{D}^\star[\rho] = 0$, then the inequality $\mathsf{D}_{N,b}(\theta)[\rho] \ge |\mathsf{S}(\theta)[\rho]|$ proved above gives $\mathsf{S}^\star[\rho] = 0$. Hence (171) holds on every coordinate. It also gives

$$\langle \overline{W}, \mathsf{S}(\theta^\star) \rangle = \nu \sum_{\rho} \mathsf{D}^\star[\rho] \overline{W}[\rho]^2 > 0. \tag{173}$$

To identify the support, write $\overline{m}_{n,c} := \langle \overline{W}, A_{n,c} \rangle$, $\mu_n := \min_{c \neq y_n} \overline{m}_{n,c}$, and $h_{n,c} := \overline{m}_{n,c} - \min\{0, \mu_n\}$. Since $\|W_t\|_{\mathrm{F}} \to \infty$ and $W_t/\|W_t\|_{\mathrm{F}} \to \overline{W}$, finite log-sum-exp asymptotics give

$$-\frac{1}{\|W_t\|_{\mathrm{F}}} \log p_{n,c}(W_t) \longrightarrow h_{n,c}.$$

Therefore every subsequential limit $\theta^\star$ is supported on the atoms minimizing $h_{n,c}$. If some $\mu_n \le 0$, then the minimum of h is zero and every minimizing atom has $\overline{m}_{n,c} \le 0$, contradicting (173). Thus every $\mu_n > 0$, so $h_{n,c} = \overline{m}_{n,c}$, proving both claims in (172). □

For fixed $\theta \in \Delta(\mathcal{Q})$, let $a_q := \operatorname{vec}(A_q)$ and define

$$\mathcal{L}_\theta(u, \lambda) := \frac{1}{2} u^\top M_{N,b}(\theta) u + \sum_{q \in \mathcal{Q}} \lambda_q (1 - a_q^\top u), \qquad g_\theta(\lambda) := \inf_{u \in \mathbb{R}^{kd}} \mathcal{L}_\theta(u, \lambda), \qquad \lambda \ge 0.$$

Lemma E.13 (Semidefinite primal-dual certificate)**.** *Fix $\theta \in \Delta(\mathcal{Q})$. Suppose U is feasible for* (167) *and $\lambda \ge 0$ satisfies*

$$M_{N,b}(\theta) \operatorname{vec}(U) = \sum_{q \in \mathcal{Q}} \lambda_q \operatorname{vec}(A_q), \qquad \lambda_q(\langle U, A_q \rangle - 1) = 0 \quad (q \in \mathcal{Q}).$$

Then $U \in \mathcal{P}_{N,b}(\theta)$, and λ is a dual optimizer. No invertibility of $M_{N,b}(\theta)$ is required.

Proof. Write $u := \mathrm{vec}(U)$, $M := M_{N,b}(\theta)$, and let V be any feasible matrix with $v := \mathrm{vec}(V)$. Since $M \succeq 0$,

$$\frac{1}{2}v^\top Mv - \frac{1}{2}u^\top Mu = \frac{1}{2}(v-u)^\top M(v-u) + \sum_q \lambda_q \langle V - U, A_q\rangle \geq 0,$$

where feasibility and complementarity are used in the last step. Thus U is primal optimal. Stationarity makes u a global minimizer of the Lagrangian for this λ, and complementarity makes the dual value equal the primal value, so λ is dual optimal by weak duality. □

Proof of Proposition E.10. Set $p^\star := \overline{W}/\gamma_{\mathrm{AP}}$, where $\overline{W} = -\overline{\Phi}$ and $\gamma_{\mathrm{AP}} > 0$ are given by Lemma E.11 and Proposition E.12. Let α be the scalar in (171) and define

$$\lambda_q^\star := \frac{\alpha}{\gamma_{\mathrm{AP}}}\theta_q^\star.$$

Then $p^\star$ is feasible because $\langle p^\star, A_q\rangle \geq 1$, and dividing (171) by γ_{AP} gives the stationarity equation in Lemma E.13. If $\lambda_q^\star > 0$, then $\theta_q^\star > 0$, so (172) gives $\langle \overline{W}, A_q\rangle = \gamma_{\mathrm{AP}}$ and hence complementarity. Lemma E.13 therefore yields $p^\star \in \mathcal{P}_{N,b}(\theta^\star)$ and dual optimality of $\lambda^\star$. Finally,

$$\|\lambda^\star\|_1 = \frac{\alpha}{\gamma_{\mathrm{AP}}}\sum_q \theta_q^\star = \frac{\alpha}{\gamma_{\mathrm{AP}}} > 0,$$

so $\theta^\star = \lambda^\star/\|\lambda^\star\|_1$. This proves the proposition. □

F Extension to Anisotropic Gaussian with Independent Orientations

This appendix proves Theorem 4.2. The proof uses clean training in an essential way: the optimizer-selected limiting directions depend only on the class means and are therefore independent of the Haar eigenbases of the test-residual covariances. We first characterize the geometry of these selected Row, Spectral, and Max directions and then compare their Gaussian test errors after exposing the independent residual orientations.

Normalization and Gaussian error transfer. Apply Lemma C.1 with the common factor $d^{-1/2}$, and suppress the rescaling in the notation. Thus throughout this appendix

$$\mu_i = \frac{1}{\sqrt{d}}\Gamma^{1/2}g_i, \qquad g_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d), \qquad X^{\mathrm{te}} \mid (Y=i) \sim \mathcal{N}\left(\mu_i, \frac{r_{\mathrm{te}}^2}{d}\Sigma_i\right),$$

and the clean training set is $X = \{(\mu_i, i)\}_{i=1}^k$. This rescaling changes neither classifier predictions nor margin-direction sets. Set $r_{\mathrm{te},0} := \sup_d r_{\mathrm{te}}(d) < \infty$. For any classifier with $w_i \neq w_j$, define

$$u_{ij}^W := \frac{w_i - w_j}{\|w_i - w_j\|_2}, \qquad Z_{i\to j}(W) := \frac{\sqrt{d}\,(u_{ij}^W)^\top \mu_i}{\sqrt{(u_{ij}^W)^\top \Sigma_i u_{ij}^W}}.$$

Lemma F.1 (Gaussian error transfer)**.** *Let $\mathcal{R}_d$ and $\mathcal{C}_d$ be nonempty classifier families. Suppose every $W \in \mathcal{R}_d$ satisfies $Z_{i\to j}(W) \geq A_d > 0$ for all $i \neq j$, while every $V \in \mathcal{C}_d$ has a directed pair $i \neq j$ satisfying $Z_{i\to j}(V) \leq B_d$, where $0 \leq B_d < A_d$. Then*

$$\frac{\sup_{W\in\mathcal{R}_d}\mathrm{Err}_{\pi,d}(W)}{\inf_{V\in\mathcal{C}_d}\mathrm{Err}_{\pi,d}(V)} \leq \frac{k-1}{\pi_{\min,d}}\exp\left(-\frac{A_d^2 - B_d^2}{2r_{\mathrm{te},0}^2}\right). \tag{174}$$

Consequently, if $A_d^2 \to \infty$, $\limsup_d B_d^2/A_d^2 < 1$, and $\log((k-1)/\pi_{\min,d}) = o(A_d^2)$, then every classifier in $\mathcal{R}_d$ has strictly larger population accuracy than every classifier in $\mathcal{C}_d$ for all sufficiently large d.

Proof. For a directed pair $i \neq j$,

$$\mathbb{P}\big(w_j^\top X^{\mathrm{te}} \geq w_i^\top X^{\mathrm{te}} \mid Y = i\big) = \overline{\Phi}\bigg(\frac{Z_{i\to j}(W)}{r_{\mathrm{te}}}\bigg),$$

where $\overline{\Phi}(x) = \mathbb{P}(G \geq x)$ for $G \sim \mathcal{N}(0,1)$. A union bound over the $k-1$ competitors gives $\mathrm{Err}_{\pi,d}(W) \leq (k-1)\overline{\Phi}(A_d/r_{\mathrm{te}})$ for every $W \in \mathcal{R}_d$. For $V \in \mathcal{C}_d$, the directed pair supplied by the assumption is contained in the multiclass error event, so $\mathrm{Err}_{\pi,d}(V) \geq \pi_{\min,d}\overline{\Phi}(B_d/r_{\mathrm{te}})$. Since $\overline{\Phi}(x)/\overline{\Phi}(y) \leq \exp(-(x^2-y^2)/2)$ for $x \geq y \geq 0$, (174) follows. The final assertion is immediate. □

For $\theta \geq 0$, write

$$T_{\theta,d} := \sum_{a=1}^{d} a^{-\theta}.$$

For every fixed $\theta \geq 0$,

$$T_{\theta,d} \asymp \begin{cases} d^{1-\theta}, & 0 \leq \theta < 1, \\ \log d, & \theta = 1, \\ 1, & \theta > 1. \end{cases} \tag{175}$$

In particular, for fixed $0 \leq s_\mu < 1$, the trace-normalized mean covariance obeys

$$c_{s_\mu} I_d \preceq \Gamma, \qquad \|\Gamma\|_{\mathrm{op}} \leq C_{s_\mu} d^{s_\mu}, \qquad \mathrm{tr}(\Gamma^2) \leq C_{s_\mu}(d + d^{2s_\mu}) \log d. \tag{176}$$

Optimizer-selected directions. Let $\mathcal{F}_{\mathrm{cl}}$ be the sigma-field generated by the class means together with the prescribed optimizer initializations and tie-breaking variables. Because training uses only $\{(\mu_i, i)\}_{i=1}^k$, $\mathcal{F}_{\mathrm{cl}}$ is independent of $O_1, \ldots, O_k$.

Lemma F.2 (Residual-exogenous optimizer limits)**.** *On the probability-one event on which the directional limits assumed in Theorem 4.2 exist, define*

$$U_{\mathrm{opt}} := \lim_{t\to\infty} \frac{W_t^{\mathrm{opt}}}{\|W_t^{\mathrm{opt}}\|_{\mathrm{F}}}, \qquad \mathrm{opt} \in \mathcal{O}_{\mathrm{row}} \cup \mathcal{O}_{\mathrm{comp}}.$$

Every U_{opt} is $\mathcal{F}_{\mathrm{cl}}$-measurable and hence independent of $O_1, \ldots, O_k$. Moreover,

$$U_{\mathrm{opt}} \in \begin{cases} \mathcal{U}^X_{\mathrm{row}}, & \mathrm{opt} \in \mathcal{O}_{\mathrm{row}}, \\ \mathcal{U}^X_{\mathrm{sp}}, & \mathrm{opt} \in \{\mathrm{SpecGD}, \mathrm{Muon}\}, \\ \mathcal{U}^X_{\max}, & \mathrm{opt} \in \{\mathrm{SignGD}, \mathrm{Signum}, \mathrm{Adam}\}, \end{cases}$$

and

$$\mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big) \longrightarrow \mathrm{Acc}_{\pi,d}(U_{\mathrm{opt}})\,. \tag{177}$$

Proof. The empirical loss, every full-batch gradient, and every optimizer state are $\mathcal{F}_{\mathrm{cl}}$-measurable, so induction gives the same measurability for every iterate and hence for every normalized directional limit. The full-batch implicit-bias results used in Theorem 3.1, together with Lemma B.2, place the three optimizer groups in the corresponding maximum-margin direction sets. Finally, population accuracy is invariant under positive rescaling, each margin direction has nonzero pairwise row differences, and the positive-definite Gaussian test laws assign zero probability to score-tie hyperplanes. Dominated convergence therefore gives (177). □

Mean geometry of the selected directions. Let $M^\mu \in \mathbb{R}^{k\times d}$ have i-th row $\mu_i^\top$, let $P := I_k - k^{-1}\mathbf{1}_k\mathbf{1}_k^\top$, and set $C^\mu := PM^\mu$. Recall $\kappa_{\mathrm{sp}}(\alpha) < 1$ from (41), and define

$$\vartheta_{\mathrm{sp}}(\alpha) := \kappa_{\mathrm{sp}}(\alpha)^2, \qquad \vartheta_{\max} := \frac{2}{\pi}, \qquad \vartheta_\alpha := \max\{\vartheta_{\mathrm{sp}}(\alpha), \vartheta_{\max}\} < 1.$$

Lemma F.3 (Selected mean geometry)**.** *Fix $0 \leq s_\mu < 1$ and any finite collection of $\mathcal{F}_{\mathrm{cl}}$-measurable Row, Spectral, and Max selectors. For every fixed $0 < \xi < 1 - s_\mu$, with probability at least $1 - e^{-d^\xi}$ for all sufficiently large d, the following hold.*

(i) Every selected Row direction satisfies

$$\min_{i\neq j}(u_{ij}^{U_{\mathrm{row}}})^\top \mu_i \geq \frac{1}{\sqrt{2}} - o(1). \tag{178}$$

(ii) For every selected U_ψ, $\psi \in \{\mathrm{sp}, \max\}$, and every fixed $\varepsilon > 0$, there exist unit vectors $a_1^\psi, \dots, a_k^\psi$, a directed edge set $\mathcal{G}_d^\psi$ with $|\mathcal{G}_d^\psi| \geq c_\varepsilon k(k-1)$, and

$$p_{ij}^\psi := \frac{a_i^\psi - a_j^\psi}{\|a_i^\psi - a_j^\psi\|_2}$$

such that

$$\max_{i\neq j} |(a_i^\psi)^\top a_j^\psi| = o(1/\log d),$$

$$\max_{(i,j)\in\mathcal{G}_d^\psi} \|u_{ij}^{U_\psi} - p_{ij}^\psi\|_2 \leq \frac{1}{\log d},$$

$$\max_{(i,j)\in\mathcal{G}_d^\psi} \big[(u_{ij}^{U_\psi})^\top \mu_i\big]_+^2 \leq \frac{\vartheta_\psi + \varepsilon}{2}.$$

The event can be chosen simultaneously for the entire finite selector collection.

Proof. Write $L_d := \log d$. Weighted Gaussian quadratic and bilinear concentration using (176), followed by a union bound over $O(k^2)$ quantities, gives

$$\max_i \big|\|\mu_i\|_2^2 - 1\big| + \max_{i\neq j} |\mu_i^\top \mu_j| \leq L_d^{-64} \tag{179}$$

except with probability at most $C\exp\{-cd^{1-s_\mu}/L_d^C\}$. This failure probability is at most e^{-d^ξ} for every fixed $\xi < 1 - s_\mu$ and sufficiently large d.

For Row, apply Proposition C.6 to the clean samples with $m_i = \mu_i$ and $e_{i,s} = \mathbf{0}$. Equation (179) gives $\max_i \|w_i - \mu_i\|_2 \leq CL_d^{-16}$, uniformly over the exact Row face. Since $\|\mu_i - \mu_j\|_2 = \sqrt{2} + o(1)$ uniformly over pairs, normalization stability yields $u_{ij}^W = (\mu_i - \mu_j)/\|\mu_i - \mu_j\|_2 + O(L_d^{-16})$, and (178) follows.

For Spectral, write the reduced singular value decomposition

$$C^\mu = U\Lambda V^\top, \qquad Q := UV^\top, \qquad r := \min\{k-1, d\},$$

and let $q_i^\top$ denote the i-th row of Q. After identifying $\mathbf{1}_k^\perp$ with $\mathbb{R}^{k-1}$, the nonzero part of C^μ has the same law as $G\Gamma^{1/2}/\sqrt{d}$, where $G \in \mathbb{R}^{(k-1)\times d}$ has i.i.d. standard Gaussian entries. Its left singular frame is therefore Haar in $\mathbf{1}_k^\perp$ and independent of the singular values. Haar projector concentration gives, on the same probability scale,

$$\max_i \left|\|q_i\|_2^2 - \frac{r}{k}\right| + \max_{i\neq j} |q_i^\top q_j| \leq L_d^{-16}. \tag{180}$$

Moreover, since $QC^{\mu\top} = U\Lambda U^\top$, conditional Haar concentration and the bound $\|\bar{\mu}\|_2 = o(1)$, where $\bar{\mu} = k^{-1}\sum_i \mu_i$, give

$$\max_{i\neq j} \left|(q_i - q_j)^\top \mu_i - \frac{\|C^\mu\|_*}{k-1}\right| \leq L_d^{-32}. \tag{181}$$

The nuclear-to-Frobenius ratio satisfies

$$\frac{\|C^\mu\|_*}{\sqrt{r}\,\|C^\mu\|_{\mathrm{F}}} \leq \kappa_{\mathrm{sp}}(\alpha) + \eta \tag{182}$$

for every fixed $\eta > 0$, with the same stretched-exponential confidence. Indeed, for fixed d, the map

$$A \mapsto \mathbb{E}_G \operatorname{tr}\left[\left(\frac{GAG^\top}{d}\right)^{1/2}\right]$$

is concave and invariant under orthogonal conjugation, hence Schur-concave in the eigenvalues of A. Among positive semidefinite A with $\operatorname{tr} A = d$, it is maximized by I_d. Gaussian concentration and the Marchenko–Pastur limit therefore bound the numerator by its isotropic limit, while (179) gives $\|C^\mu\|_{\mathrm{F}}/\sqrt{k} = 1 + o(1)$.

It remains to transfer the polar geometry to every exact Spectral margin direction. Since $s_\mu < 1$, (175) gives $\Gamma \succeq c_{s_\mu} I_d$. Hence the positive singular values of C^μ dominate $\sqrt{c_{s_\mu}}$ times those of the corresponding standard Gaussian matrix. If $\alpha \neq 1$, the smallest positive singular value is therefore bounded below by a constant with probability $1 - e^{-cd}$. If $\alpha = 1$, the standard Gaussian hard-edge estimate and a dyadic union bound give

$$\frac{1}{r}\#\{\ell : \lambda_\ell(C^\mu) \le t\} \le Ct, \qquad L_d^{-8} \le t \le 1. \tag{183}$$

with probability at least $1 - Ce^{-cd/L_d^{16}}$.

Let $W \in \mathcal{U}_{\mathrm{sp}}^X$, put $B := PW/\|PW\|_{\mathrm{op}}$, and note that $\|PW\|_{\mathrm{op}} = \|W\|_{\mathrm{op}}$ because centering preserves all margins. The polar factor Q has clean margin $\|C^\mu\|_*/(k-1) + O(L_d^{-32})$ by (181). Summing the normalized margin constraints for B therefore gives

$$\langle B, C^\mu \rangle \ge \|C^\mu\|_* - kL_d^{-32}.$$

Writing $d_\ell := 1 - u_\ell^\top B v_\ell \in [0, 2]$, where u_ℓ, v_ℓ are the singular vectors of C^μ, yields

$$\frac{1}{r}\sum_{\ell=1}^{r} \lambda_\ell(C^\mu) d_\ell \le CL_d^{-32}.$$

The positive spectral edge for $\alpha \neq 1$, and (183) with a split at $t = L_d^{-8}$ for $\alpha = 1$, imply

$$\frac{1}{r}\|B - Q\|_{\mathrm{F}}^2 \le CL_d^{-8}.$$

The centered pair identity then gives

$$\sup_{W \in \mathcal{U}_{\mathrm{sp}}^X} \frac{1}{k(k-1)} \sum_{i \neq j} \left\| u_{ij}^W - \frac{q_i - q_j}{\|q_i - q_j\|_2} \right\|_2^2 \le CL_d^{-8}. \tag{184}$$

Define

$$a_i^{\mathrm{sp}} := \frac{q_i}{\|q_i\|_2}, \qquad p_{ij}^{\mathrm{sp}} := \frac{a_i^{\mathrm{sp}} - a_j^{\mathrm{sp}}}{\|a_i^{\mathrm{sp}} - a_j^{\mathrm{sp}}\|_2}.$$

Equation (180) gives $\max_{i \neq j} |(a_i^{\mathrm{sp}})^\top a_j^{\mathrm{sp}}| \le CL_d^{-16}$ and shows that p_{ij}^{sp} differs from $(q_i - q_j)/\|q_i - q_j\|_2$ by $O(L_d^{-16})$. Combining (181), (182), and $\|C^\mu\|_{\mathrm{F}}/\sqrt{k} = 1 + o(1)$ gives

$$\max_{i \neq j} \big[(p_{ij}^{\mathrm{sp}})^\top \mu_i\big]_+ \le \frac{\kappa_{\mathrm{sp}}(\alpha) + 2\eta}{\sqrt{2}}.$$

Markov's inequality applied to (184) shows that all but $o(k^2)$ ordered pairs satisfy $\|u_{ij}^{U_{\mathrm{sp}}} - p_{ij}^{\mathrm{sp}}\|_2 \le L_d^{-1}$. Choosing $\eta > 0$ sufficiently small in terms of the prescribed ε proves part (ii) for Spectral selectors.

For Max, use the common Haar eigenbasis of the mean covariance. Write

$$\mu_i = U_\mu \widetilde{\mu}_i, \qquad \widetilde{\mu}_i := \frac{D_{s_\mu}^{1/2} g_i}{\sqrt{T_{s_\mu, d}}}, \qquad \widetilde{c}_i := \widetilde{\mu}_i - \frac{1}{k}\sum_{\ell=1}^{k} \widetilde{\mu}_\ell.$$

Then the i-th row of C^μ is $(U_\mu \widetilde{c}_i)^\top$. Weighted Gaussian concentration gives

$$\max_i \big|\|\widetilde{c}_i\|_2^2 - 1\big| + \max_{i \neq j} |\widetilde{c}_i^\top \widetilde{c}_j| + \max_i \big|\widetilde{c}_i^\top \widetilde{\mu}_i - 1\big| + \max_{i \neq j} |\widetilde{c}_j^\top \widetilde{\mu}_i| \le L_d^{-64} \tag{185}$$

outside an event of probability $C \exp\{-cd^{1-s_\mu}/L_d^C\}$.

Put $S^\mu := \operatorname{sign}(C^\mu)$. Conditional on the vectors in (185), standard smoothed-sign Haar concentration, using the identities

$$\mathbb{E}[\operatorname{sign}(q^\top x)\operatorname{sign}(q^\top y)] = \frac{2}{\pi}\arcsin\left(\frac{x^\top y}{\|x\|_2\|y\|_2}\right)$$

and

$$\sqrt{d}\,\mathbb{E}[\operatorname{sign}(q^\top x)\,q^\top v] = \sqrt{d}\,\mathbb{E}|q_1|\frac{x^\top v}{\|x\|_2} = \left(\sqrt{\frac{2}{\pi}} + o(1)\right)\frac{x^\top v}{\|x\|_2},$$

gives

$$\left|\frac{\|C^\mu\|_{\mathrm{sum}}}{k\sqrt{d}} - \sqrt{\frac{2}{\pi}}\right| \le CL_d^{-32}, \tag{186}$$

$$\max_{i\neq j}\left|\frac{(S_i^\mu - S_j^\mu)^\top \mu_i}{\sqrt{d}} - \sqrt{\frac{2}{\pi}}\right| \le CL_d^{-32}, \tag{187}$$

$$\max_{i\neq j}\left|\frac{(S_i^\mu)^\top S_j^\mu}{d}\right| \le CL_d^{-8}, \tag{188}$$

$$\frac{1}{kd}\#\{(i,a) : \sqrt{d}\,|C^\mu[i,a]| \le L_d^{-16}\} \le CL_d^{-16}. \tag{189}$$

For example, these bounds follow by replacing the sign function by a piecewise-linear approximation at scale L_d^{-128}, applying concentration on $\mathsf{O}(d)$, and then controlling the smoothing error by the spherical small-ball probability. The resulting failure probability is $C\exp\{-cd/L_d^C\}$.

Let $W \in \mathcal{U}_{\max}^X$ and $B := W/\|W\|_{\max}$. The sign matrix is a feasible unit-Max direction with margin at least $(\sqrt{2/\pi} - CL_d^{-32})\sqrt{d}$, so $\gamma_{\max}(X)$ is at least this value. Summing the normalized margin constraints for B and using (186) gives

$$\frac{\|C^\mu\|_{\mathrm{sum}} - \langle B, C^\mu\rangle}{k\sqrt{d}} \le CL_d^{-32}.$$

Since $|B[i,a]| \le 1$, splitting the coordinates at $\sqrt{d}\,|C^\mu[i,a]| = L_d^{-16}$ and using (189) yields

$$\sup_{W\in\mathcal{U}_{\max}^X}\frac{1}{kd}\left\|\frac{W}{\|W\|_{\max}} - S^\mu\right\|_{\mathrm{F}}^2 \le CL_d^{-16}. \tag{190}$$

Equation (188) implies $\|S_i^\mu - S_j^\mu\|_2^2 = 2d(1 + O(L_d^{-8}))$. The centered pair identity and (190) therefore give

$$\sup_{W\in\mathcal{U}_{\max}^X}\frac{1}{k(k-1)}\sum_{i\neq j}\left\|u_{ij}^W - \frac{S_i^\mu - S_j^\mu}{\|S_i^\mu - S_j^\mu\|_2}\right\|_2^2 \le CL_d^{-16}. \tag{191}$$

Define

$$a_i^{\max} := \frac{S_i^\mu}{\sqrt{d}}, \qquad p_{ij}^{\max} := \frac{S_i^\mu - S_j^\mu}{\|S_i^\mu - S_j^\mu\|_2}.$$

Equations (187)–(188) give

$$\max_{i\neq j}|(a_i^{\max})^\top a_j^{\max}| \le CL_d^{-8}, \qquad \max_{i\neq j}\left|\left[(p_{ij}^{\max})^\top \mu_i\right]^2 - \frac{1}{\pi}\right| \le CL_d^{-8}.$$

Markov's inequality applied to (191) leaves all but $o(k^2)$ pairs with direction error at most L_d^{-1}. Since $1/\pi = \vartheta_{\max}/2$, part (ii) follows for Max selectors. The preceding bounds are uniform over the complete exact Spectral and Max faces, so the same event works for every member of the fixed finite selector collection. □

Residual geometry. The preceding lemma depends only on the class means. We now expose the independent test-residual bases.

Lemma F.4 (Residual energies for selected directions)**.** *Condition on an event from Lemma F.3, and consider the corresponding finite collection of selected Row, Spectral, and Max directions.*

(i) If $0 \le s_\Sigma < 1$, then for every fixed $0 < \chi < 1 - s_\Sigma$, with conditional probability at least $1 - e^{-d^\chi}$ for all sufficiently large d,

$$\max_{\text{selected } U} \max_{i \neq j} \left| (u_{ij}^U)^\top \Sigma_i u_{ij}^U - 1 \right| = o(1). \tag{192}$$

(ii) Suppose $s_\Sigma \ge 1$ and put

$$c_d := \frac{\log d}{T_{s_\Sigma, d}}.$$

For every fixed $b > 0$ and $\varepsilon > 0$, all selected Row directions satisfy

$$\max_{i \neq j} (u_{ij}^{U_{\mathrm{row}}})^\top \Sigma_i u_{ij}^{U_{\mathrm{row}}} \le 1 + (4 + 2b + \varepsilon) c_d \tag{193}$$

except with conditional probability at most $C_{b,\varepsilon} d^{-b}$. For every fixed $q \in (0,1)$, each selected competitor U_ψ, $\psi \in \{\mathrm{sp}, \max\}$, has an edge $(i,j) \in \mathcal{G}_d^\psi$ such that

$$(u_{ij}^{U_\psi})^\top \Sigma_i u_{ij}^{U_\psi} \ge 4 q c_d + \mathbf{1}_{\{s_\Sigma = 1\}} (1 - o(1)) \tag{194}$$

except with conditional probability at most $C_{q,a} d^{-a}$ for every fixed $0 < a < 1 - q$.

Proof. For a unit vector u independent of O_i, write $O_i^\top u = g / \|g\|_2$. The quadratic form is

$$u^\top \Sigma_i u = \frac{d}{\|g\|_2^2 T_{s_\Sigma, d}} \sum_{r=1}^{d} r^{-s_\Sigma} g_r^2.$$

When $s_\Sigma < 1$, weighted chi-square concentration gives deviations $o(1)$ with failure at most e^{-d^χ} for every $\chi < 1 - s_\Sigma$. A union bound over the finite number of selectors and their $O(k^2) = O(d^2)$ pair directions proves (192).

Now suppose $s_\Sigma \ge 1$. Laurent–Massart's inequality gives, for every $t > 0$,

$$\frac{1}{T_{s_\Sigma, d}} \sum_{r=1}^{d} r^{-s_\Sigma} g_r^2 \le 1 + \frac{2\sqrt{T_{2s_\Sigma, d} t} + 2t}{T_{s_\Sigma, d}}$$

except with probability e^{-t}, while $\|g\|_2^2 / d = 1 + o(1)$ exponentially fast. Taking $t = (2 + b) \log d$ and union-bounding over $O(d^2)$ selected Row pair directions gives (193); the square-root term is $o(c_d)$.

For the competitor lower bound, fix $q \in (0,1)$ and $0 < a < 1 - q$, and choose $q_+ \in (q, 1 - a)$. Write $L_d := \log d$, let $h_i := O_i e_1$, and represent $h_i = \xi_i / \|\xi_i\|_2$ using independent $\xi_i \sim \mathcal{N}(\mathbf{0}, I_d)$. Fix one selected competitor and its code (a_i^ψ) and edge set $\mathcal{G}_d^\psi$ from Lemma F.3. Put $t_d := \sqrt{2 q_+ L_d}$ and define

$$N_d := \sum_{(i,j) \in \mathcal{G}_d^\psi} \mathbf{1}\{\xi_i^\top a_i^\psi \ge t_d,\ \xi_i^\top a_j^\psi \le -t_d\}.$$

The coherence bound $\max_{i \neq j} |(a_i^\psi)^\top a_j^\psi| = o(1/L_d)$ implies, uniformly over the relevant distinct indices,

$$\mathbb{P}(\xi_i^\top a_i^\psi \ge t_d, \xi_i^\top a_j^\psi \le -t_d) = d^{-2q_+ + o(1)}, \qquad \mathbb{P}(\xi_i^\top a_i^\psi \ge t_d, \xi_i^\top a_j^\psi \le -t_d, \xi_i^\top a_\ell^\psi \le -t_d) = d^{-3q_+ + o(1)}.$$

Indeed, the two- and three-dimensional covariance matrices differ from the identity by $o(1/L_d)$, and Gaussian density comparison on the relevant orthants changes the logarithmic tail asymptotics by $o(L_d)$. Edge indicators with different first indices are independent. Since $|\mathcal{G}_d^\psi| = \Theta(d^2)$,

$$\mathbb{E} N_d = d^{2 - 2q_+ + o(1)}, \qquad \mathrm{Var}(N_d) \le \mathbb{E} N_d + d^{3 - 3q_+ + o(1)}.$$

Chebyshev's inequality gives

$$\mathbb{P}(N_d = 0) \le d^{-(1-q_+)+o(1)} = o(d^{-a}). \tag{195}$$

On $\{N_d > 0\}$, choose a counted edge using a fixed ordering of $\mathcal{G}_d^\psi$. This choice depends only on the class-side information and $(\xi_i)_{i=1}^k$, not on the remaining residual eigenframes. Uniform Gaussian norm concentration and $\|a_i^\psi - a_j^\psi\|_2 = \sqrt{2} + o(1)$ give

$$h_i^\top p_{ij}^\psi \ge \frac{2t_d}{(1+o(1))\sqrt{2d}}.$$

For any fixed $A > 0$, spherical concentration and a union bound give, except with probability $O(d^{-A})$,

$$\max_{(i,j)\in\mathcal{G}_d^\psi} \sqrt{d}\,\left|h_i^\top (u_{ij}^{U_\psi} - p_{ij}^\psi)\right| \le C_A L_d^{-1/2}, \qquad \max_{i\ne j} d\left|h_i^\top u_{ij}^{U_\psi}\right|^2 \le C_A L_d.$$

Here the first bound uses $\|u_{ij}^{U_\psi} - p_{ij}^\psi\|_2 \le L_d^{-1}$, and all directions in these bounds are fixed after conditioning on the class-side information. The selected edge therefore satisfies

$$d\big(h_i^\top u_{ij}^{U_\psi}\big)^2 \ge (4q_+ - o(1))L_d \ge 4qL_d$$

for all sufficiently large d. Since the top eigenvalue of Σ_i is $d/T_{s_\Sigma,d}$, its contribution is at least $4qc_d$.

Suppose $s_\Sigma = 1$. Conditional on all top eigenvectors, their auxiliary Gaussian radii, and the selected edge, the remaining eigenframe of O_i is Haar on $h_i^\perp$. Write $a_{ij} := h_i^\top u_{ij}^{U_\psi}$. The remaining contribution has the conditional distribution

$$\frac{d(1-a_{ij}^2)}{T_{1,d}} \frac{\sum_{r=2}^d r^{-1} g_r^2}{\sum_{r=2}^d g_r^2}, \qquad g_2, \dots, g_d \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

The preceding upper bound gives $a_{ij}^2 = O(L_d/d)$. The weighted chi-square lower-tail inequality, with tail parameter $L_d^{3/2}$, gives

$$\sum_{r=2}^d r^{-1} g_r^2 \ge T_{1,d} - 1 - O(L_d^{3/4})$$

except with probability $\exp(-L_d^{3/2})$, since $\sum_{r=2}^d r^{-2} = O(1)$. Also, $\sum_{r=2}^d g_r^2 = (d-1)(1 + O(L_d^{-2}))$ except with probability $C\exp(-cd/L_d^4)$. Thus the remaining contribution is at least $1 - O(L_d^{-1/4}) = 1 - o(1)$ with superpolynomial confidence. Choosing $A > a$, integrating these conditional bounds, and using (195) proves (194) with failure at most $C_{q,a} d^{-a}$. A finite union handles all selected competitors. □

Selected-direction accuracy ordering.

Proposition F.5 (Selected-direction ordering)**.** *Fix $0 \le s_\mu < 1$, $s_\Sigma \ge 0$, and any finite collection of $\mathcal{F}_{\text{cl}}$-measurable Row, Spectral, and Max selectors. Assume*

$$\log \frac{k-1}{\pi_{\min,d}} = o(d/\log d).$$

Then the following hold.

(i) If $0 \le s_\Sigma < 1$, then for every fixed $0 < \chi < \min\{1 - s_\mu, 1 - s_\Sigma\}$, every selected Row classifier has strictly larger population accuracy than every selected Spectral or Max classifier with probability at least $1 - e^{-d^\chi}$ for every sufficiently large d.

(ii) If $s_\Sigma \ge 1$, then for every fixed

$$0 < \zeta < \frac{2(1-\vartheta_\alpha)}{2+\vartheta_\alpha}, \tag{196}$$

the same strict ordering holds with probability at least $1 - C_\zeta d^{-\zeta}$ for every sufficiently large d.

Proof. Suppose first that $s_\Sigma < 1$. Fix $0 < \chi < \min\{1 - s_\mu, 1 - s_\Sigma\}$, choose $\chi_+ \in (\chi, \min\{1 - s_\mu, 1 - s_\Sigma\})$, and fix $0 < \varepsilon_0 < 1 - \vartheta_\alpha$. Applying Lemma F.3 and Lemma F.4(i) with confidence exponent χ_+ implies that every selected Row direction has

$$Z_{i\to j}(U_{\mathrm{row}})^2 \geq (1/2 - o(1))d$$

uniformly over all directed pairs, whereas every selected competitor U_ψ has a directed pair with

$$[Z_{i\to j}(U_\psi)]_+^2 \leq \frac{\vartheta_\psi + \varepsilon_0}{2}\, d + o(d).$$

Since $\vartheta_\psi + \varepsilon_0 < 1$, the competitor-to-Row squared-signal ratio has limsup strictly below one. The Row coefficient is of order d, and $\log((k-1)/\pi_{\min,d}) = o(d/\log d) = o(d)$, so Lemma F.1 gives the strict accuracy ordering. The two mean and residual events have joint failure at most e^{-d^χ} after increasing the dimension threshold.

Now suppose $s_\Sigma \geq 1$. Fix ζ satisfying (196), choose $b > \zeta$ sufficiently close to ζ, and then choose q so that

$$\vartheta_\alpha \left(1 + \frac{b}{2}\right) < q < 1 - \zeta. \tag{197}$$

Such a choice is possible because of (196). Choose fixed $\varepsilon_R, \varepsilon_M > 0$ sufficiently small that

$$(\vartheta_\alpha + \varepsilon_M)(4 + 2b + \varepsilon_R) < 4q, \qquad (\vartheta_\alpha + \varepsilon_M)(5 + 2b + \varepsilon_R) < 1 + 4q.$$

The first inequality holds with zero slacks by (197); the second follows from the same condition and $\vartheta_\alpha < 1$. Apply Lemma F.3 with signal slack ε_M and Lemma F.4 with Row-energy slack ε_R. They supply squared comparison amplitudes satisfying

$$A_{R,d}^2 \geq \frac{(1/2 - o(1))d}{1 + (4 + 2b + \varepsilon_R)c_d},$$
$$B_{\psi,d}^2 \leq \frac{((\vartheta_\psi + \varepsilon_M)/2)d}{4qc_d + \mathbf{1}_{\{s_\Sigma = 1\}}(1 - o(1))}.$$

If $s_\Sigma > 1$, then $c_d \to \infty$, and

$$\limsup_d \frac{B_{\psi,d}^2}{A_{R,d}^2} \leq \frac{(\vartheta_\psi + \varepsilon_M)(4 + 2b + \varepsilon_R)}{4q} < 1.$$

If $s_\Sigma = 1$, then $c_d \to 1$, and

$$\limsup_d \frac{B_{\psi,d}^2}{A_{R,d}^2} \leq \frac{(\vartheta_\psi + \varepsilon_M)(5 + 2b + \varepsilon_R)}{1 + 4q} < 1.$$

Moreover, $A_{R,d}^2 \gtrsim d/\log d$, so the prior assumption is sufficient for Lemma F.1. The Row residual-energy failure is $O(d^{-b})$, while the competitor-extreme failure is $O(d^{-a})$ for every $a < 1 - q$. Choose $a \in (\zeta, 1 - q)$. The mean-geometry failure is stretched-exponential, so intersecting the finitely many selected-direction events gives total failure at most $C_\zeta d^{-\zeta}$. □

Proof of Theorem 4.2. Lemma F.2 turns the seven optimizer limits into a finite collection of $\mathcal{F}_{\mathrm{cl}}$-measurable Row, Spectral, and Max selectors and gives convergence of their population accuracies. Choose any fixed

$$0 < c < \frac{2(1 - \vartheta_\alpha)}{2 + \vartheta_\alpha}.$$

If $0 \leq s_\Sigma < 1$, choose any fixed $0 < \chi < \min\{1 - s_\mu, 1 - s_\Sigma\}$. Proposition F.5(i) gives failure probability at most e^{-d^χ}, which is at most d^{-c} for every sufficiently large d. If $s_\Sigma \geq 1$, Proposition F.5(ii), applied with $\zeta = c$, gives failure probability at most $C_c d^{-c}$. After enlarging a constant C, both cases therefore hold with probability at least $1 - Cd^{-c}$. The exponent c and the constant C can be chosen independently of s_μ and s_Σ; only the sufficiently-large-d threshold depends on these exponents. On this event, every selected Row limit has strictly larger population accuracy than every selected Spectral or Max limit. Equation (177) then gives

$$\min_{\mathrm{opt}\in\mathcal{O}_{\mathrm{row}}} \liminf_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big) > \max_{\mathrm{opt}\in\mathcal{O}_{\mathrm{comp}}} \limsup_{t\to\infty} \mathrm{Acc}_{\pi,d}\big(W_t^{\mathrm{opt}}\big),$$

which is (11). □

G COORDINATE-SPECIFIC ROWGDM–ADAM REVERSAL

This section proves the coordinate-aligned transition and the recovery under a common Haar rotation in Proposition 4.5. The comparisons hold uniformly over the complete Row and Max maximum-margin solution sets.

Setup. We rescale all features by $d^{-1/2}$, which preserves the maximum-margin direction sets and population accuracies by Lemma C.1. Let $k/d \to \alpha \in (0,\infty)$, and suppose

$$\Gamma_d = \mathrm{diag}(\gamma_{1,d},\dots,\gamma_{d,d}), \qquad \frac{1}{d}\operatorname{tr}\Gamma_d = 1, \qquad 0 < c_\Gamma \le \gamma_{a,d} \le C_\Gamma < \infty.$$

The independent class means and clean training set are

$$\mu_i = \frac{1}{\sqrt{d}}\Gamma_d^{1/2} g_i, \qquad g_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d), \qquad \mathcal{X} = \{(\mu_i, i) : i \in [k]\}.$$

Write $M \in \mathbb{R}^{k\times d}$ for the matrix with rows $\mu_i^\top$, and define

$$\mathcal{F}(M) := \{W : (w_i - w_j)^\top \mu_i \ge 1 \text{ for every } i \ne j\}.$$

The two minimum-norm solution sets are

$$\mathcal{M}_{\mathrm{row}}(M) := \operatorname*{arg\,min}_{W\in\mathcal{F}(M)} \|W\|_{\mathrm{row}}, \qquad \mathcal{M}_{\max}(M) := \operatorname*{arg\,min}_{W\in\mathcal{F}(M)} \|W\|_{\max}.$$

For $\Sigma \succ 0$ and a classifier with positive directed mean margins, set

$$L_{i\to j}(W;\Sigma) := \frac{((w_i - w_j)^\top \mu_i)^2}{(w_i - w_j)^\top \Sigma (w_i - w_j)}, \qquad E_d(W;\Sigma) := \min_{i\ne j} L_{i\to j}(W;\Sigma).$$

The test distribution is

$$\mathbb{P}(Y = i) = \pi_{i,d}, \qquad X^{\mathrm{te}} \mid (Y = i) \sim \mathcal{N}\left(\mu_i, \frac{r_{\mathrm{te}}^2}{d}\Sigma\right),$$

where $0 < r_{\mathrm{te}} \le r_{\mathrm{te},0} < \infty$, $\pi_{\min,d} := \min_i \pi_{i,d} > 0$, and

$$\log(1/\pi_{\min,d}) = o(d). \tag{198}$$

Write $\mathrm{Err}_{\pi,d} = 1 - \mathrm{Acc}_{\pi,d}$ and $g_0 := \sqrt{2/\pi}$. Constants denoted by c, C may change between displays; their permitted dependence is specified in each statement.

Lemma G.1 (Gaussian coefficient comparison)**.** *Every classifier with positive directed mean margins satisfies*

$$\pi_{\min,d}\,\overline{\Phi}\left(\frac{\sqrt{dE_d(W;\Sigma)}}{r_{\mathrm{te}}}\right) \le \mathrm{Err}_{\pi,d}(W) \le (k-1)\,\overline{\Phi}\left(\frac{\sqrt{dE_d(W;\Sigma)}}{r_{\mathrm{te}}}\right), \tag{199}$$

where $\overline{\Phi}(x) = \mathbb{P}(Z \ge x)$ for $Z \sim \mathcal{N}(0,1)$. If nonempty classifier families $\mathcal{R}, \mathcal{C}$ have positive directed mean margins and satisfy $E_d(R;\Sigma) \ge A > B \ge E_d(V;\Sigma) > 0$ for every $R \in \mathcal{R}$ and $V \in \mathcal{C}$, then

$$\frac{\sup_{R\in\mathcal{R}} \mathrm{Err}_{\pi,d}(R)}{\inf_{V\in\mathcal{C}} \mathrm{Err}_{\pi,d}(V)} \le \frac{k-1}{\pi_{\min,d}} \exp\left(-\frac{d(A-B)}{2r_{\mathrm{te},0}^2}\right). \tag{200}$$

In particular, a fixed positive coefficient gap gives strict uniform accuracy ordering for all sufficiently large d, without any additional probabilistic loss.

Proof. The directed pair error equals $\overline{\Phi}(\sqrt{dL_{i\to j}(W;\Sigma)}/r_{\mathrm{te}})$. A classwise union bound gives the upper bound in (199), and a pair attaining E_d gives the lower bound. For $x \ge y \ge 0$, changing variables in the Gaussian tail integral gives

$$\overline{\Phi}(x) = \int_y^\infty \frac{e^{-(u+x-y)^2/2}}{\sqrt{2\pi}}\,du \le e^{-(x^2-y^2)/2}\overline{\Phi}(y).$$

Applying this inequality with $x = \sqrt{dA}/r_{\mathrm{te}}$ and $y = \sqrt{dB}/r_{\mathrm{te}}$ proves (200). The final assertion follows from $k = \Theta(d)$ and (198). □

Lemma G.2 (Deterministic whole-face stability). *Fix $h_* > 0$ and $K < \infty$. There exist $\varepsilon_0, C > 0$, depending only on h_*, K, with the following property. Let $0 < \varepsilon \le \varepsilon_0$, $k^{-1} \le \varepsilon$, and $h \in [h_*, 1]$. Suppose deterministic means have no zero coordinates, and, with $\bar\mu = k^{-1}\sum_i \mu_i$ and $T_i = \operatorname{sign}(\mu_i)$, satisfy*

$$\max_i |\|\mu_i\|_2^2 - 1| \le \varepsilon, \qquad \max_{i\neq j} |\mu_i^\top \mu_j| \le \varepsilon, \qquad \|\bar\mu\|_2 \le \varepsilon,$$

$$\max_i \left| \frac{\|\mu_i\|_1}{\sqrt{d}} - h \right| \le \varepsilon, \qquad \max_{i\neq j} \frac{|T_j^\top \mu_i|}{\sqrt{d}} \le \varepsilon, \qquad \max_{i\neq j} \frac{|T_i^\top T_j|}{d} \le \varepsilon.$$

Assume also that, for $s \in \{\varepsilon^{1/2}, \varepsilon^{1/8}\}$,

$$\max_i \frac{1}{d} \#\{a : \sqrt{d}\,|\mu_i[a]| \le s\} \le Ks + \varepsilon.$$

Then both solution sets are nonempty, and, writing $t_{\max} := \min_{W\in\mathcal{F}(M)} \|W\|_{\max}$,

$$\sup_{W\in\mathcal{M}_{\mathrm{row}}(M)} \max_i \|w_i - \mu_i\|_2 \le C\varepsilon^{1/4}, \tag{201}$$

$$\left| \frac{1}{t_{\max}\sqrt{d}} - h \right| + \sup_{W\in\mathcal{M}_{\max}(M)} \max_i \frac{\|w_i/t_{\max} - T_i\|_2}{\sqrt{d}} \le C\varepsilon^{1/16}. \tag{202}$$

Proof. The matrix with rows $\mu_i/(1-2\varepsilon)$ is feasible, so the optimal row radius R_* is at most $\sqrt{1+\varepsilon}/(1-2\varepsilon) \le 1 + C\varepsilon$. For any feasible W, averaging the class-i constraints over competitors gives

$$(w_i - \bar w)^\top \mu_i \ge 1 - \frac{1}{k}, \qquad \bar w := \frac{1}{k}\sum_i w_i. \tag{203}$$

For a Row minimizer, Cauchy–Schwarz and the row variance identity therefore imply

$$\frac{(1-1/k)^2}{1+\varepsilon} \le \frac{1}{k}\sum_i \|w_i - \bar w\|_2^2 \le R_*^2 - \|\bar w\|_2^2.$$

Thus $\|\bar w\|_2 \le C\varepsilon^{1/2}$, and (203) gives $w_i^\top \mu_i \ge 1 - C\varepsilon^{1/2}$ uniformly in i. Expanding $\|w_i - \mu_i\|_2^2$ proves (201).

The sign matrix divided by $\sqrt{d}(h-2\varepsilon)$ is feasible, so $t_{\max} \le [\sqrt{d}(h-2\varepsilon)]^{-1}$. Fix a Max minimizer and set $H = W/t_{\max}$, $\bar H = k^{-1}\sum_i H_i$, and

$$D_i := \frac{\|\mu_i\|_1 - H_i^\top \mu_i}{\sqrt{d}} \ge 0.$$

Since $\|H\|_{\max} = 1$, averaging (203) over i yields

$$\begin{aligned} \frac{1}{k}\sum_i D_i &\le h + \varepsilon - \frac{1-1/k}{t_{\max}\sqrt{d}} - \frac{\bar H^\top \bar\mu}{\sqrt{d}} \\ &\le h + \varepsilon - (1-1/k)(h-2\varepsilon) + \|\bar\mu\|_2 \le C\varepsilon. \end{aligned}$$

For every $s > 0$, the coordinate bound on H gives

$$\frac{\|H_i - T_i\|_2^2}{d} \le \frac{2D_i}{s} + \frac{4}{d}\#\{a : \sqrt{d}\,|\mu_i[a]| \le s\}. \tag{204}$$

Indeed, outside the indicated small-coordinate set, use $(H_i[a] - T_i[a])^2 \le 2(1 - H_i[a]T_i[a])$, and use the bound four on that set. Taking $s = \varepsilon^{1/2}$ and averaging gives $\|H - T\|_{\mathrm{F}}^2/(kd) \le C\varepsilon^{1/2}$. Moreover, $\|k^{-1}\sum_i T_i\|_2^2/d \le k^{-1} + \varepsilon \le 2\varepsilon$, so Jensen's inequality gives $\|\bar H\|_2/\sqrt{d} \le C\varepsilon^{1/4}$. Returning to the individual constraints now yields

$$D_i \le h + \varepsilon - (1-1/k)(h-2\varepsilon) + \frac{\|\bar H\|_2 \|\mu_i\|_2}{\sqrt{d}} \le C\varepsilon^{1/4}.$$

Taking $s = \varepsilon^{1/8}$ in (204) proves the claimed uniform row error for H. Finally, averaging the constraints and using $H_i^\top \mu_i \le \|\mu_i\|_1$ gives

$$h - 2\varepsilon \le \frac{1}{t_{\max}\sqrt{d}} \le \frac{h + 2\varepsilon}{1 - 1/k} = h + O(\varepsilon).$$

This proves (202). The feasible witnesses and compact norm sublevel sets ensure existence of both minima. □

Lemma G.3 (Quantitative whole-face geometry in aligned coordinates)**.** *Define*

$$\bar{\gamma}_{1/2,d} := \frac{1}{d}\sum_a \sqrt{\gamma_{a,d}}, \qquad h_d := g_0 \bar{\gamma}_{1/2,d}, \qquad \bar{\gamma}_{2,d} := \frac{1}{d}\sum_a \gamma_{a,d}^2.$$

There exist $\varepsilon_0, c, C > 0$, depending only on c_Γ, C_Γ, such that, whenever $0 < \varepsilon \le \varepsilon_0$ and $k \ge C\varepsilon^{-2}$, an event $\mathcal{A}_d(\varepsilon)$ satisfies

$$\mathbb{P}(\mathcal{A}_d(\varepsilon)^c) \le Ck^2 e^{-cd\varepsilon^2}. \tag{205}$$

On this event, both solution sets are nonempty and

$$\sup_{W \in \mathcal{M}_{\mathrm{row}}(M)} \max_i \|w_i - \mu_i\|_2 \le C\varepsilon^{1/4}, \tag{206}$$

$$\sup_{W \in \mathcal{M}_{\max}(M)} \max_i \frac{\|w_i/t_{\max} - T_i\|_2}{\sqrt{d}} \le C\varepsilon^{1/16}, \qquad T_i := \mathrm{sign}(\mu_i). \tag{207}$$

Also, uniformly over $i \ne j$,

$$\begin{aligned} \big|\|\mu_i\|_2^2 - 1\big| + |\mu_i^\top \mu_j| &\le C\varepsilon, \\ \left|\frac{(T_i - T_j)^\top \mu_i}{\sqrt{d}} - h_d\right| + \left|\frac{\|T_i - T_j\|_2^2}{2d} - 1\right| &\le C\varepsilon. \end{aligned} \tag{208}$$

For $\Sigma_\lambda := (1-\lambda)I_d + \lambda\Gamma_d$ and $A_d(\lambda) := 1 + \lambda(\bar{\gamma}_{2,d} - 1)$, the same event satisfies

$$\begin{aligned} \sup_{\lambda \in [0,1]} \max_{i \ne j} \big|(\mu_i - \mu_j)^\top \Sigma_\lambda (\mu_i - \mu_j) - 2A_d(\lambda)\big| &\le C\varepsilon, \\ \sup_{\lambda \in [0,1]} \max_{i \ne j} \left|\frac{(T_i - T_j)^\top \Sigma_\lambda (T_i - T_j)}{2d} - 1\right| &\le C\varepsilon. \end{aligned} \tag{209}$$

Proof. We verify the inputs of Lemma G.2 at precision ε. The independent-coordinate Gaussian quadratic-form and Bernstein bounds used below follow, for example, from the concentration tools in Vershynin (2012). For the quadratic terms, the identity $\mathbb{E}e^{u(Z^2-1)} = e^{-u}(1-2u)^{-1/2}$ for $|u| < 1/2$, together with the bounded spectral weights, gives deviation probability at most $2e^{-cd\varepsilon^2}$ for each normalized quadratic form. Polarization gives the corresponding bilinear bounds. The variables $\sqrt{\gamma_{a,d}}|g_i[a]|$ have uniformly bounded exponential moments, so Bernstein's inequality gives

$$\mathbb{P}\left(\left|\frac{\|\mu_i\|_1}{\sqrt{d}} - h_d\right| > \varepsilon\right) \le 2e^{-cd\varepsilon^2}.$$

For $j \ne i$, conditioning on T_j gives $T_j^\top \mu_i/\sqrt{d} \sim \mathcal{N}(0, 1/d)$. The products $T_i[a]T_j[a]$ are independent Rademacher variables across a, so their averages satisfy the same exponential bound. For either prescribed threshold s, Gaussian small-ball probabilities and Hoeffding's inequality give

$$\mathbb{P}\left(\frac{1}{d}\#\{a : \sqrt{d}\,|\mu_i[a]| \le s\} > Ks + \varepsilon\right) \le e^{-2d\varepsilon^2},$$

where K depends only on c_Γ. Also, $\bar{\mu}$ has covariance $\Gamma_d/(kd)$; when $k \ge C\varepsilon^{-2}$, the event $\|\bar{\mu}\|_2 > \varepsilon$ has probability at most Ce^{-cd}. All coordinates are nonzero almost surely, and $g_0\sqrt{c_\Gamma} \le h_d \le g_0 < 1$. A union bound over these $O(k^2)$ events verifies the deterministic lemma's hypotheses and proves (206)–(208).

For (209), it suffices to treat $\lambda = 0, 1$, since the quadratic forms and their asserted centers are affine in λ. The Gaussian difference $\mu_i - \mu_j$ has covariance $2\Gamma_d/d$, giving mean quadratic form

$2d^{-1}\operatorname{tr}(\Gamma_d\Sigma_\lambda) = 2A_d(\lambda)$. The same quadratic-form bound applies at both endpoints. For the sign contrasts, $(T_i[a] - T_j[a])^2$ are independent across coordinates, lie in $\{0, 4\}$, and have mean two. Their bounded weighted sums therefore concentrate about $2\operatorname{tr}\Sigma_\lambda = 2d$ with failure at most $2e^{-cd\varepsilon^2}$ at each endpoint. Adding these events proves (205) and (209). □

Proposition G.4 (An exact Row–Max transition under aligned test anisotropy)**.** *Suppose*

$$\bar\gamma_{1/2,d} \to \bar\gamma_{1/2}, \qquad \bar\gamma_{2,d} \to \bar\gamma_2, \tag{210}$$

and set

$$h := g_0\bar\gamma_{1/2}, \qquad h^2\bar\gamma_2 > 1.$$

For

$$\Sigma_\lambda := (1-\lambda)I_d + \lambda\Gamma_d, \qquad \lambda \in [0,1],$$

there are constants $\varepsilon_0, C > 0$, depending only on the spectral bounds, such that, whenever $0 < \varepsilon \le \varepsilon_0$ and $k \ge C\varepsilon^{-2}$, the event $\mathcal{A}_d(\varepsilon)$ in Lemma G.3 satisfies

$$\sup_{\lambda\in[0,1]}\sup_{W\in\mathcal{M}_{\mathrm{row}}(M)}\left|E_d(W;\Sigma_\lambda) - \frac{1}{2A_d(\lambda)}\right| \le C\varepsilon^{1/4},$$
$$\sup_{\lambda\in[0,1]}\sup_{W\in\mathcal{M}_{\max}(M)}\left|E_d(W;\Sigma_\lambda) - \frac{h_d^2}{2}\right| \le C\varepsilon^{1/16}. \tag{211}$$

The limiting coefficients cross uniquely at

$$\lambda_c := \frac{h^{-2}-1}{\bar\gamma_2 - 1} \in (0,1).$$

For every fixed $\lambda \ne \lambda_c$, there exist $c_\lambda, C_\lambda > 0$, independent of d, such that, for every sufficiently large d,

$$\lambda < \lambda_c \implies \mathbb{P}\left\{\inf_{W\in\mathcal{M}_{\mathrm{row}}}\mathrm{Acc}_{\pi,d}(W) > \sup_{W\in\mathcal{M}_{\max}}\mathrm{Acc}_{\pi,d}(W)\right\} \ge 1 - C_\lambda e^{-c_\lambda d},$$
$$\lambda > \lambda_c \implies \mathbb{P}\left\{\inf_{W\in\mathcal{M}_{\max}}\mathrm{Acc}_{\pi,d}(W) > \sup_{W\in\mathcal{M}_{\mathrm{row}}}\mathrm{Acc}_{\pi,d}(W)\right\} \ge 1 - C_\lambda e^{-c_\lambda d}.$$

At $\lambda = \lambda_c$, only equality of the limiting coefficients is asserted.

Proof. The spectral bounds give $c_\Gamma I_d \preceq \Sigma_\lambda \preceq C_\Gamma I_d$ for every $\lambda \in [0,1]$, since trace normalization implies $c_\Gamma \le 1 \le C_\Gamma$. For Row, (206) allows us to replace $w_i - w_j$ by $\mu_i - \mu_j$ with Euclidean error $C\varepsilon^{1/4}$. The latter vector has directed signal $1 + O(\varepsilon)$ and test-noise quadratic form $2A_d(\lambda) + O(\varepsilon)$, uniformly over pairs and λ. The denominators are bounded away from zero, proving the first estimate in (211). For Max, scale invariance allows us to use the contrast $(H_i - H_j)/\sqrt{d}$, where $H = W/t_{\max}$. By (207), this contrast differs from $(T_i - T_j)/\sqrt{d}$ by at most $C\varepsilon^{1/16}$. The canonical sign contrast has directed signal $h_d + O(\varepsilon)$ and test-noise quadratic form $2 + O(\varepsilon)$, proving the second estimate. These arguments also prove the same estimates for every individual $L_{i\to j}$.

Since $\bar\gamma_{1/2} \le 1$, we have $0 < h < 1$; the condition $h^2\bar\gamma_2 > 1$ implies $\bar\gamma_2 > 1$ and $0 < \lambda_c < 1$. Fix $\lambda \ne \lambda_c$, and let

$$\Delta_\lambda := \left|\frac{1}{2[1+\lambda(\bar\gamma_2 - 1)]} - \frac{h^2}{2}\right| > 0.$$

Choose a fixed $\varepsilon > 0$ small enough that the sum of the two errors in (211) is at most $\Delta_\lambda/4$. By (210), the finite-dimensional coefficient gap has the asserted sign and magnitude at least $3\Delta_\lambda/4$ for all sufficiently large d. Thus the complete-face gap is at least $\Delta_\lambda/2$ on $\mathcal{A}_d(\varepsilon)$. Lemma G.1 proves the corresponding strict accuracy ordering. Finally, $k = \Theta(d)$ allows the factor k^2 in (205) to be absorbed into $C_\lambda e^{-c_\lambda d}$. No convergence rate for either spectral summary is required; their convergence affects only the sufficiently-large-d threshold. □

The Haar-rotation statements below hold equally for Haar measure on $O(d)$ and on $\mathrm{SO}(d)$. Indeed, for the fixed orthonormal frames used below, both measures induce the same uniform Stiefel distribution, and, for sufficiently large d, conditional on a fixed two-frame, the unexposed direction is uniform on the orthogonal sphere in either case.

Lemma G.5 (Quantitative sign statistics under a Haar rotation)**.** *Let deterministic vectors $x_1, \dots, x_k \in \mathbb{R}^d$ satisfy $1/2 \le \|x_i\|_2 \le 2$, let R be Haar on $O(d)$ or $\mathrm{SO}(d)$, and put $T_i = \mathrm{sign}(Rx_i)$. There are universal $\varepsilon_0, c, C > 0$ such that, for $0 < \varepsilon \le \varepsilon_0$ and any two prescribed thresholds $s_1, s_2 \in (0, 1]$, an event of probability at least $1 - Ck^2 e^{-cd\varepsilon^6}$ satisfies*

$$\max_{i,j} \left| \frac{T_i^\top R x_j}{\sqrt{d}} - g_0 \frac{x_i^\top x_j}{\|x_i\|_2} \right| \le C\varepsilon,$$
$$\max_{i,j} \left| \frac{T_i^\top T_j}{d} - \frac{2}{\pi} \arcsin\left(\frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2} \right) \right| \le C\varepsilon,$$
$$\max_i \frac{1}{d} \#\{a : \sqrt{d}\,|(Rx_i)[a]| \le s_\ell\} \le C s_\ell + C\varepsilon^2, \qquad \ell = 1, 2. \tag{212}$$

Proof. Fix a pair and choose an orthonormal basis $F \in \mathbb{R}^{d \times m}$ of its span, where $m \le 2$. Write $x_i = Fu_i$, $x_j = Fu_j$, and let $Z \in \mathbb{R}^{d \times m}$ have independent standard Gaussian entries. Then

$$\sqrt{d}\, RF \overset{\mathrm{d}}{=} Y := ZB, \qquad B := (Z^\top Z/d)^{-1/2}.$$

Indeed, $Y/\sqrt{d}$ has orthonormal columns and its distribution is invariant under every left orthogonal transformation. The Gaussian quadratic-form bound gives

$$\mathbb{P}\big(\|Z^\top Z/d - I_m\|_{\mathrm{op}} > \varepsilon^3\big) \le Ce^{-cd\varepsilon^6}.$$

On the complementary event, $\|B - I_m\|_{\mathrm{op}} \le C\varepsilon^3$, and hence

$$\frac{1}{d}\|(Y - Z)u\|_2^2 \le C\varepsilon^6$$

for either coefficient vector u. Set $X = Zu$, $\widetilde{X} = Yu$, and $b = \varepsilon^2$. Gaussian small-ball probabilities and Hoeffding's inequality imply $d^{-1}\#\{a : |X_a| \le b\} \le C\varepsilon^2$, except with probability $Ce^{-cd\varepsilon^4}$. Consequently,

$$\frac{1}{d}\#\{a : \mathrm{sign}(\widetilde{X}_a) \ne \mathrm{sign}(X_a)\} \le \frac{1}{d}\#\{a : |X_a| \le b\} + \frac{\|\widetilde{X} - X\|_2^2}{db^2} \le C\varepsilon^2.$$

Thus the normalized Euclidean distance between the corresponding sign vectors is at most $C\varepsilon$. By Cauchy–Schwarz, replacing Y by Z changes the empirical averages involving one sign factor and one linear factor by at most $C\varepsilon$, while the empirical correlations between the corresponding sign vectors change by at most $C\varepsilon^2$. For a standard Gaussian row $z^\top$ of Z, the corresponding expectations are

$$\mathbb{E}[\mathrm{sign}(z^\top u) z^\top v] = g_0 \frac{u^\top v}{\|u\|_2}, \qquad \mathbb{E}[\mathrm{sign}(z^\top u)\,\mathrm{sign}(z^\top v)] = \frac{2}{\pi} \arcsin\left(\frac{u^\top v}{\|u\|_2 \|v\|_2} \right).$$

The first identity follows by conditioning $z^\top v$ on $z^\top u$, and the second is the standard arcsine formula for the signs of two jointly Gaussian variables. Bernstein's and Hoeffding's inequalities concentrate the corresponding empirical averages within ε with failure at most $Ce^{-cd\varepsilon^2}$. Finally, for either prescribed s,

$$\frac{1}{d}\#\{a : |\widetilde{X}_a| \le s\} \le \frac{1}{d}\#\{a : |X_a| \le s + b\} + \frac{\|\widetilde{X} - X\|_2^2}{db^2} \le Cs + C\varepsilon^2$$

except with probability $Ce^{-cd\varepsilon^4}$. A union bound over the pairs and the two thresholds proves (212); independence between different pairs is not needed. □

Lemma G.6 (Generic common rotation restores the Row coefficient advantage)**.** *For deterministic means, define*

$$\delta_d := \max_i |\|\mu_i\|_2^2 - 1| + \max_{i \ne j} |\mu_i^\top \mu_j|.$$

Let $\Sigma \succ 0$ satisfy $\mathrm{tr}\,\Sigma = d$ and $c_\Sigma I_d \preceq \Sigma \preceq C_\Sigma I_d$, and set

$$\mu_i^\circ = R\mu_i, \qquad \Sigma^\circ = R\Sigma R^\top, \qquad v_{ij} := \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|_2}, \qquad a_{ij} := v_{ij}^\top \Sigma v_{ij},$$

where R is an independent Haar matrix. There are $\varepsilon_0, c, C > 0$, depending only on c_Σ, C_Σ, such that, whenever

$$0 < \varepsilon \le \varepsilon_0, \qquad \delta_d + d^{-1} + k^{-1} \le \varepsilon^2,$$

an event of probability at least $1 - Ck^2 e^{-cd\varepsilon^6}$, over R alone, satisfies

$$\sup_{W \in \mathcal{M}^\circ_{\mathrm{row}}} \max_{i \neq j} \left| L_{i\to j}(W; \Sigma^\circ) - \frac{1}{2a_{ij}} \right| \le C\varepsilon^{1/4},$$

$$\sup_{W \in \mathcal{M}^\circ_{\max}} \max_{i \neq j} \left| L_{i\to j}(W; \Sigma^\circ) - \frac{1}{2a_{ij} + \pi - 2} \right| \le C\varepsilon^{1/16}. \tag{213}$$

Here the superscript $\circ$ denotes solution sets and accuracies for the rotated problem. Consequently, if $\delta_d \to 0$, $k/d \to \alpha \in (0, \infty)$, and (198) *holds, there exist $c_*, C_* > 0$, independent of d, such that, for every sufficiently large d,*

$$\mathbb{P}_R \left\{ \inf_{W \in \mathcal{M}^\circ_{\mathrm{row}}} \mathrm{Acc}^\circ_{\pi,d}(W) > \sup_{W \in \mathcal{M}^\circ_{\max}} \mathrm{Acc}^\circ_{\pi,d}(W) \right\} \ge 1 - C_* e^{-c_* d}. \tag{214}$$

Proof. The hypotheses imply $\|\bar{\mu}\|_2^2 \le (1 + \delta_d)/k + \delta_d \le C\varepsilon^2$. Apply Lemma G.5 to the means. The norm and coherence assumptions, together with (212), give, for $T_i = \mathrm{sign}(R\mu_i)$,

$$\max_i \left| \frac{\|R\mu_i\|_1}{\sqrt{d}} - g_0 \right| + \max_{i \neq j} \frac{|T_j^\top R\mu_i|}{\sqrt{d}} + \max_{i \neq j} \frac{|T_i^\top T_j|}{d} \le C\varepsilon.$$

Choose a fixed sufficiently large constant C_0 and apply that lemma with thresholds $(C_0\varepsilon)^{1/2}$ and $(C_0\varepsilon)^{1/8}$. After decreasing ε_0, Lemma G.2 applies at precision $C_0\varepsilon$, with $h = g_0$. It follows that every Row minimizer is within $C\varepsilon^{1/4}$ per row of the rotated means, while every rescaled Max minimizer is within $C\varepsilon^{1/16}\sqrt{d}$ per row of T_i.

It remains to calculate the test-noise energy of the canonical sign contrast

$$b_{ij} := \frac{T_i - T_j}{\|T_i - T_j\|_2}.$$

Fix a pair, let $S = \mathrm{span}\{\mu_i, \mu_j\}$, and choose an orthonormal basis $F \in \mathbb{R}^{d\times 2}$ of S. The Gram matrix of the pair is uniformly well-conditioned under the assumed bound on δ_d. Conditioning only on RF, the vector b_{ij} is fixed whenever it is defined. Set $p = (RF)^\top b_{ij}$, $q = Fp$, and $r^2 = 1 - \|p\|_2^2$. The unexposed action of R on the orthogonal complement is Haar, so conditionally

$$R^\top b_{ij} = q + r\zeta, \tag{215}$$

where ζ is uniform on the unit sphere of $S^\perp$. On the sign-statistics event, (212) gives

$$b_{ij}^\top R\mu_i = \frac{g_0}{\sqrt{2}} + O(\varepsilon), \qquad b_{ij}^\top R\mu_j = -\frac{g_0}{\sqrt{2}} + O(\varepsilon).$$

The same two inner products for $g_0 v_{ij}$ are $g_0/\sqrt{2} + O(\delta_d)$ and $-g_0/\sqrt{2} + O(\delta_d)$. Inverting the well-conditioned two-by-two Gram matrix therefore yields

$$\|q - g_0 v_{ij}\|_2 \le C\varepsilon, \qquad |r^2 - (1 - g_0^2)| \le C\varepsilon. \tag{216}$$

For the conditional complement calculation, let P_S denote the orthogonal projector onto S. Representing ζ by a normalized standard Gaussian in $S^\perp$ and applying the bounded-weight quadratic-form and Gaussian linear-form bounds gives

$$\left| \zeta^\top \Sigma \zeta - \frac{\mathrm{tr}((I_d - P_S)\Sigma)}{d - 2} \right| + |q^\top \Sigma \zeta| \le C\varepsilon$$

with conditional failure at most $Ce^{-cd\varepsilon^2}$, uniformly in the exposed frame. The trace ratio is $1 + O(d^{-1})$, since $\mathrm{tr}\,\Sigma = d$ and $\mathrm{rank}\, P_S = 2$. The conditional failure bound can be integrated and then union-bounded over pairs; we do not condition on the sign vectors of the other classes. Combining this bound with (215) and (216) gives

$$\max_{i \neq j} \left| b_{ij}^\top \Sigma^\circ b_{ij} - (g_0^2 a_{ij} + 1 - g_0^2) \right| \le C\varepsilon.$$

The canonical Row direction Rv_{ij} has directed signal $1/\sqrt{2} + O(\delta_d)$ and test-noise energy exactly a_{ij}. The canonical Max direction has directed signal $g_0/\sqrt{2} + O(\varepsilon)$, and

$$\frac{g_0^2/2}{g_0^2 a_{ij} + 1 - g_0^2} = \frac{1}{2a_{ij} + \pi - 2}.$$

Both canonical contrast norms are bounded away from zero after the Max contrasts are divided by $\sqrt{d}$. Normalization stability, the whole-face bounds already proved, and $\Sigma^\circ \succeq c_\Sigma I_d$ therefore transfer these coefficients uniformly to the complete solution sets, proving (213). The combined failure is at most $Ck^2 e^{-cd\varepsilon^6}$.

To pass to the minimum directed coefficient, put $a_* := \max_{i \neq j} a_{ij}$. Then (213) implies

$$\inf_{W \in \mathcal{M}^\circ_{\mathrm{row}}} E_d(W; \Sigma^\circ) - \sup_{W \in \mathcal{M}^\circ_{\max}} E_d(W; \Sigma^\circ) \geq \frac{1}{2a_*} - \frac{1}{2a_* + \pi - 2} - C\varepsilon^{1/16}.$$

Since $a_* \leq C_\Sigma$, the difference of the two displayed fractions is at least

$$\Delta_{\mathrm{rot}} := \frac{\pi - 2}{2C_\Sigma(2C_\Sigma + \pi - 2)} > 0. \tag{217}$$

Choose a fixed $\varepsilon > 0$ so that $C\varepsilon^{1/16} \leq \Delta_{\mathrm{rot}}/2$. The regularity assumptions eventually ensure $\delta_d + d^{-1} + k^{-1} \leq \varepsilon^2$. Lemma G.1 gives the strict accuracy ordering, and absorbing k^2 into the exponential proves (214). □

Proposition G.7 (Opposite Row–Max orderings before and after a common Haar rotation)**.** *Under the hypotheses of Proposition G.4, fix $\lambda \in (\lambda_c, 1]$. Compare the original problem with means μ_i and covariance Σ_λ to the problem obtained by the independent common rotation $\mu_i \mapsto R\mu_i$, $\Sigma_\lambda \mapsto R\Sigma_\lambda R^\top$. There exist $c, C > 0$, independent of d, such that, for every sufficiently large d, with joint probability at least $1 - Ce^{-cd}$,*

$$\begin{aligned} \inf_{W \in \mathcal{M}_{\max}} \mathrm{Acc}_{\pi,d}(W) &> \sup_{W \in \mathcal{M}_{\mathrm{row}}} \mathrm{Acc}_{\pi,d}(W), \\ \inf_{W \in \mathcal{M}^\circ_{\mathrm{row}}} \mathrm{Acc}^\circ_{\pi,d}(W) &> \sup_{W \in \mathcal{M}^\circ_{\max}} \mathrm{Acc}^\circ_{\pi,d}(W). \end{aligned} \tag{218}$$

The two problems have identical class Gram matrices, class-mean norms and pair distances, mean singular values, test-noise eigenvalues, and mean-contrast test-noise variances.

Proof. Proposition G.4 gives the first ordering with exponentially small failure. For the second ordering, choose the fixed precision ε in the last paragraph of Lemma G.6. Gaussian quadratic and bilinear concentration gives

$$\mathbb{P}\{\delta_d > \varepsilon^2/2\} \leq Ck^2 e^{-cd\varepsilon^4}.$$

For all sufficiently large d, we also have $d^{-1} + k^{-1} \leq \varepsilon^2/2$. Conditional on any means in this event, Lemma G.6 applies with the same constants, since $c_\Gamma I_d \preceq \Sigma_\lambda \preceq C_\Gamma I_d$. Its conditional failure is at most $Ck^2 e^{-cd\varepsilon^6}$. Integrating this uniform conditional bound and intersecting with the first ordering proves (218). The invariances follow directly from orthogonality; in particular,

$$(Rv_{ij})^\top (R\Sigma_\lambda R^\top)(Rv_{ij}) = v_{ij}^\top \Sigma_\lambda v_{ij}.$$

□

Proof of Proposition 4.5. Lemma C.1 identifies the direction sets and accuracies in the main-text normalization with those analyzed here. Since $h = g_0 \bar{\gamma}_{1/2}$, the transition point in Proposition G.4 is

$$\frac{h^{-2} - 1}{\bar{\gamma}_2 - 1} = \frac{\pi/(2\bar{\gamma}_{1/2}^2) - 1}{\bar{\gamma}_2 - 1} = \lambda_c.$$

That proposition therefore gives the diagonal complete-face ordering with probability at least $1 - C_\lambda e^{-c_\lambda d}$ for every fixed $\lambda \neq \lambda_c$ and sufficiently large d.

For every fixed $\lambda \in [0, 1]$, the conditional argument in the proof of Proposition G.7 gives the rotated Row-over-Max ordering with probability at least $1 - C_{\mathrm{rot}} e^{-c_{\mathrm{rot}} d}$. This argument does not require

$\lambda > \lambda_c$. Its spectral bounds and the positive gap in (217) are uniform in λ, so $c_{\mathrm{rot}}, C_{\mathrm{rot}}$ can be chosen independently of λ.

At each fixed dimension on these events, Proposition B.14 and Theorem B.7, together with Lemma B.2, place the normalized RowGDM and Adam trajectories asymptotically in their respective Row and Max maximum-margin direction sets. Every classifier in these direction sets has nonzero pairwise row differences. The compactness and continuity argument of Lemma C.21 applies to the present positive-definite Gaussian test covariances as well, since all score-tie hyperplanes have probability zero. It transfers each uniform complete-face ordering to the stated strict comparison between the trajectory liminf and limsup accuracies, without an additional exceptional event or a directional-convergence assumption. The same argument applies to RowGD. □

Corollary G.8 (Spectral dispersion implies a transition)**.** *Suppose the assumptions of Proposition 4.5 concerning the existence of the limiting spectral summaries hold. Assume additionally that, for some fixed* $0 < \underline{\gamma} < 1 < \overline{\gamma} < \infty$,

$$\underline{\gamma} \le \gamma_{a,d} \le \overline{\gamma} \qquad \text{for all } a, d,$$

and that

$$v_\Gamma := \lim_{d\to\infty} \frac{1}{d}\sum_{a=1}^{d}(\gamma_{a,d}-1)^2 > \frac{\pi}{2q_*^2} - 1, \qquad q_* := \frac{(\overline{\gamma}-1)\sqrt{\underline{\gamma}} + (1-\underline{\gamma})\sqrt{\overline{\gamma}}}{\overline{\gamma}-\underline{\gamma}}.$$

Then $2\bar{\gamma}_{1/2}^2\bar{\gamma}_2/\pi > 1$*, and hence the transition point in Proposition 4.5 lies in* $(0,1)$*. In particular, the transition condition does not require a special spectral shape: within a fixed bounded spectral range, the displayed lower bound on the limiting spectral variance is sufficient.*

Proof. Concavity of $x \mapsto \sqrt{x}$ places its graph above the chord joining $(\underline{\gamma}, \sqrt{\underline{\gamma}})$ and $(\overline{\gamma}, \sqrt{\overline{\gamma}})$. Averaging this chord inequality and using $d^{-1}\sum_a \gamma_{a,d} = 1$ gives $\bar{\gamma}_{1/2} \ge q_*$. Trace normalization also gives $\bar{\gamma}_2 = 1 + v_\Gamma$, and therefore

$$\frac{2\bar{\gamma}_{1/2}^2\bar{\gamma}_2}{\pi} \ge \frac{2q_*^2(1+v_\Gamma)}{\pi} > 1.$$

□

Remark G.9 (Non-vacuity of the dispersion condition)**.** *Write* $l := \underline{\gamma}$ *and* $u := \overline{\gamma}$*. The variance permitted by the spectral range is itself bounded. Indeed, since* $l \le \gamma_{a,d} \le u$*, we have* $(\gamma_{a,d}-l)(u-\gamma_{a,d}) \ge 0$*, or equivalently* $\gamma_{a,d}^2 \le (l+u)\gamma_{a,d} - lu$*. Averaging and using the trace normalization* $d^{-1}\sum_a \gamma_{a,d} = 1$ *gives* $\bar{\gamma}_2 \le l + u - lu$*, and hence* $v_\Gamma = \bar{\gamma}_2 - 1 \le (1-l)(u-1)$*. This bound is sharp: it is attained asymptotically by a two-point spectral profile supported on* l *and* u *with respective masses* $(u-1)/(u-l)$ *and* $(1-l)/(u-l)$*.*

Consequently, the dispersion condition in Corollary G.8 is non-vacuous if and only if $(1-l)(u-1) > \pi/(2q_*^2) - 1$*. Using* $q_* = (1+\sqrt{lu})/(\sqrt{l}+\sqrt{u})$*, this is equivalently*

$$\frac{2(1+\sqrt{lu})^2(l+u-lu)}{\pi(\sqrt{l}+\sqrt{u})^2} > 1.$$

For each fixed $l \in (0,1)$*, the left-hand side is strictly increasing in* $u > 1$*, so there is a unique critical value* $u_c(l) > 1$ *above which the sufficient dispersion condition can be satisfied.*

H Additional Experiments and Experimental Details

H.1 Synthetic Experiments: Setup and Additional Results

Setup. The main-text experiment in Figure 2 (left) uses $k = 60$ classes and dimension $d = 20$. For each class i, we draw $\mu_i \sim \mathcal{N}(\mathbf{0}, I_d/d)$, draw the number of training examples independently and uniformly from $\{1, \dots, 5\}$, and generate $x_{i,r}^{\mathrm{tr}} = \mu_i + 0.2236\,\varepsilon_{i,r}^{\mathrm{tr}}$ and $x_{i,r}^{\mathrm{te}} = \mu_i + 2.2361\,\varepsilon_{i,r}^{\mathrm{te}}$, where all residual vectors are independent $\mathcal{N}(\mathbf{0}, I_d/d)$. Test-class proportions are drawn independently from $\mathrm{Dirichlet}(\mathbf{1}_k)$, and each of the 30 paired seeds uses 120,000 evaluation examples. All methods within a seed use the same training and evaluation data.

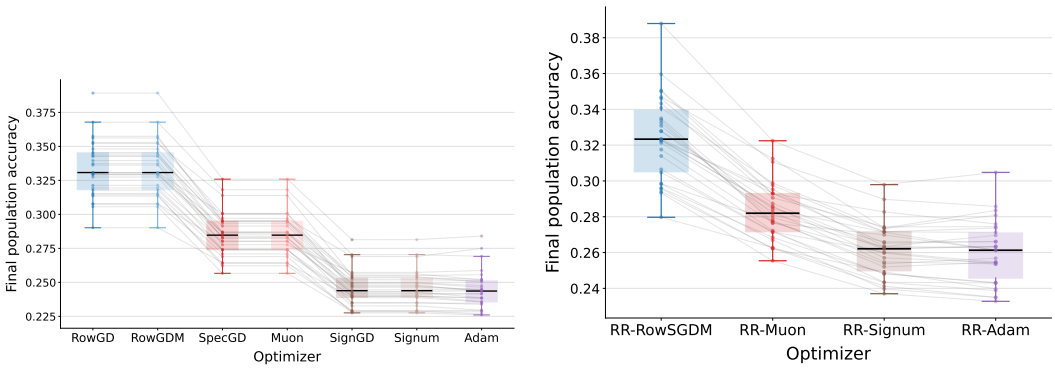

Figure 3: **Additional optimizer comparisons on the baseline synthetic model. Left:** The deterministic comparison in Figure 2 (left), augmented with Signum. **Right:** Random-reshuffling comparison of RR-RowSGDM, RR-Muon, RR-Signum, and RR-Adam with batch size $b = 1$. Points are paired seeds and gray lines connect methods evaluated on the same dataset.

Optimization and learning-rate selection. We train a bias-free multiclass linear classifier from $W_0 = \mathbf{0}$ with cross-entropy for 20,000 updates in float64 and use $\eta_t = \eta_0(t+1)^{-1/2}$. For every optimizer and seed, we sweep the common effective-Frobenius coefficients $c \in \{0.03, 0.1, 0.3, 1, 3, 10, 30\}$, setting $\eta_0 = c/\|\Delta_0\|_{\mathrm{F}}$, and select c by maximum final evaluation accuracy, with ties broken by lower training loss and then smaller c. Because the same evaluation sample is used for learning-rate selection and reporting, these results are comparisons under the stated discrete sweep rather than untouched-test estimates; some RowGD and RowGDM runs select the lower endpoint of the grid.

Additional deterministic and stochastic comparisons. Figure 3 complements the six-optimizer deterministic comparison in Figure 2 (left) in two ways. The left panel adds Signum to give the full deterministic comparison among RowGD, RowGDM, SpecGD, Muon, SignGD, Signum, and Adam, while the right panel compares the corresponding random-reshuffling methods RR-RowSGDM, RR-Muon, RR-Signum, and RR-Adam. RowGDM, Muon, and Signum use momentum $\beta = 0.9$; SpecGD and Muon use the exact compact-SVD polar factor; and Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$ without bias correction or a numerical-stability constant. For random reshuffling, we use batch size $b = 1$, draw an independent permutation of the finite training set at each epoch, use momentum $\beta = 0.99$ for RR-RowSGDM, RR-Muon, and RR-Signum, and use $(\beta_1, \beta_2) = (0.99, 0.999)$ for RR-Adam. The deterministic panel preserves the Row-over-Spectral and Row-over-coordinate-wise ordering of the main-text experiment, and RR-RowSGDM likewise outperforms all three random-reshuffling comparators on every paired seed. RR-Adam is included as an empirical comparator; our theoretical stochastic Adam comparison is formulated through AdamProxy.

Auxiliary geometries. Figure 4 supplements the main-text optimizer comparison with three reference methods that are not part of our theoretical ordering. Frobenius-normalized GD uses $\Delta_t = G_t/\|G_t\|_{\mathrm{F}}$, ColumnGD independently normalizes each column of G_t in Euclidean norm, and ordinary GD uses $\Delta_t = G_t$. The figure uses **FrobGD** as an abbreviation for Frobenius-normalized GD. These methods provide additional finite-dimensional context rather than additional theoretical comparators.

Robustness of the deterministic ordering. To test whether the ordering in Figure 2 (left) depends on the particular baseline configuration, we repeat the same six-optimizer comparison under eleven one-factor-at-a-time variations. We vary the training-noise coefficient from 0.2236 to 0.0447 or 0.4472, the test-noise coefficient from 2.2361 to 1.3416 or 3.1305, the test-class proportions from $\mathrm{Dirichlet}(\mathbf{1}_k)$ to balanced classes or $\mathrm{Dirichlet}(0.3\mathbf{1}_k)$, the number of training examples per class to exactly one or five, the problem size to $(k, d) = (30, 20)$ or $(120, 40)$, and the class-mean distribution from Gaussian to uniform on the unit sphere. In all Gaussian-noise settings, the residual vectors are drawn from $\mathcal{N}(\mathbf{0}, I_d/d)$; in the $(120, 40)$ setting, the training- and test-noise coefficients are 0.3162 and 3.1623, respectively. Each setting uses 30 paired seeds and the same training and learning-rate

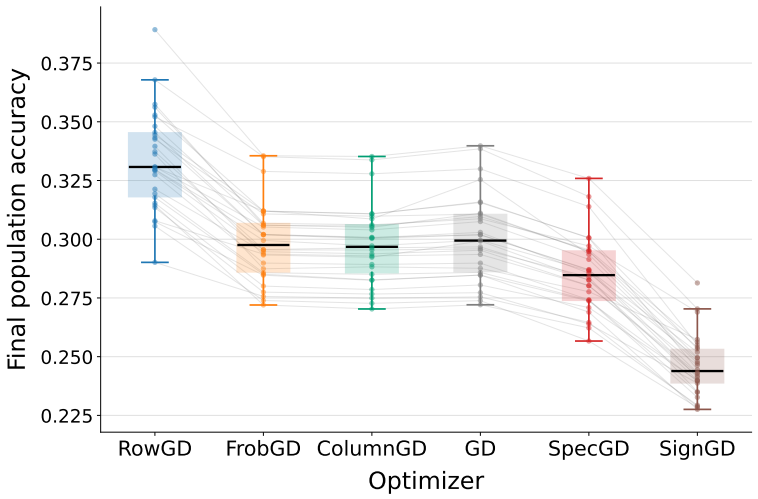

Figure 4: **Auxiliary optimization geometries.** RowGD, FrobGD, ColumnGD, ordinary GD, SpecGD, and SignGD on the same 30 paired baseline datasets, where FrobGD denotes Frobenius-normalized GD.

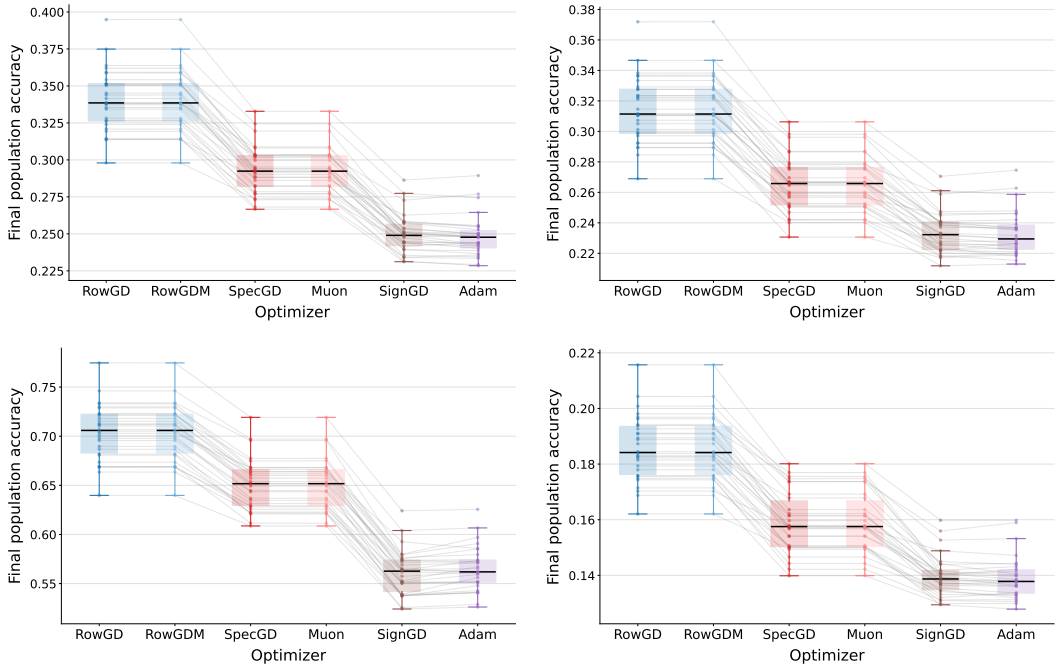

Figure 5: **Robustness to training and test noise.** Top: training-noise coefficient 0.0447 (left) and 0.4472 (right), compared with 0.2236 in Figure 2 (left). Bottom: test-noise coefficient 1.3416 (left) and 3.1305 (right), compared with 2.2361 in the main-text setting. In all cases, the residual vectors are drawn from $\mathcal{N}(\mathbf{0}, I_d/d)$.

protocol as the main-text experiment. Across all settings, RowGD and RowGDM have the two highest mean accuracies; each outperforms SignGD and Adam on every paired seed, and each outperforms SpecGD and Muon on all 30 seeds except in the one- and five-training-example settings, where the Row methods win 29 of 30 comparisons. Figures 5–7 show that the main-text ordering is stable across these changes.

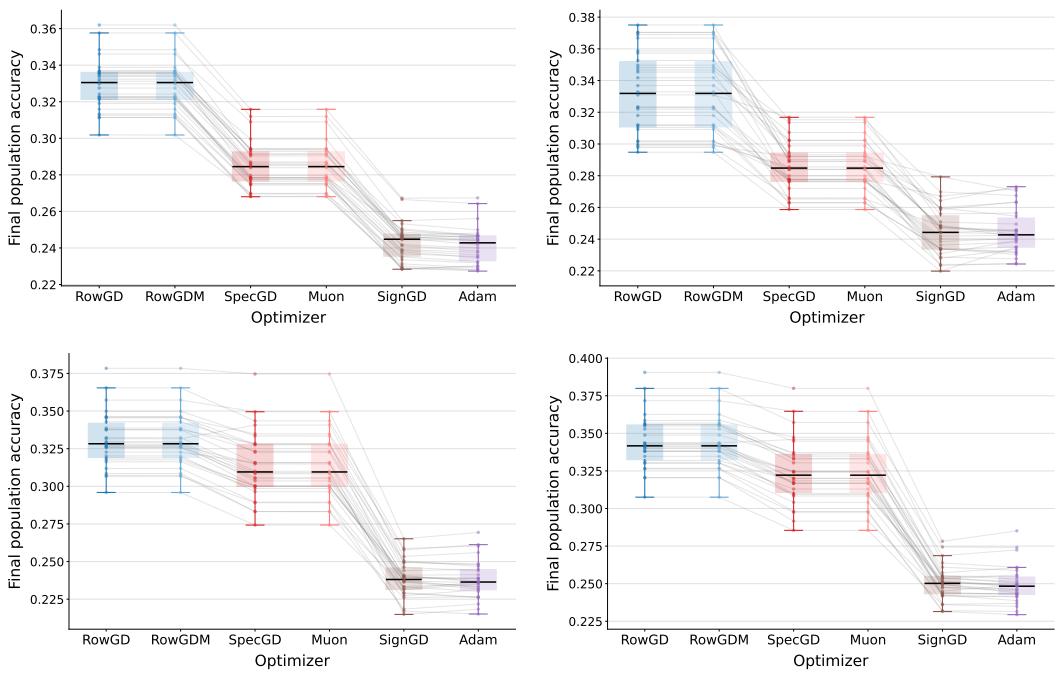

Figure 6: **Robustness to class proportions and training sample count.** Top: balanced test classes (left) and $\mathrm{Dirichlet}(0.3\mathbf{1}_k)$ test proportions (right), compared with $\mathrm{Dirichlet}(\mathbf{1}_k)$ in Figure 2 (left). Bottom: exactly one (left) and five (right) training examples per class, compared with independent counts drawn uniformly from $\{1, \ldots, 5\}$ in the main-text setting.

Anisotropic synthetic experiments. We also test the anisotropic predictions of Section 4. Unless otherwise noted, the training and evaluation protocol is the same as in the baseline synthetic experiment above. Figure 8 shows a representative independent-orientation setting with power-law exponents $s_\mu = 0.5$ and $s_\Sigma = 2$; across six such settings, RowGD and RowGDM outperform SpecGD, Muon, SignGD, and Adam on all $30/30$ paired seeds in every setting, with conservative mean advantages between 2.82 and 5.73 percentage points. Figure 9 illustrates the coordinate-aligned reversal and recovery phenomenon from Proposition 4.5: in the diagonal setting, RowGDM outperforms Adam on all $30/30$ seeds at $\lambda = 0.25$, Adam outperforms RowGDM on all $30/30$ seeds at $\lambda = 1$, and applying the same Haar rotation restores the RowGDM-over-Adam ordering on all $30/30$ seeds. The coordinate-aligned reversal-and-recovery experiments use 10^6 training steps to reduce sensitivity to finite-time transients.

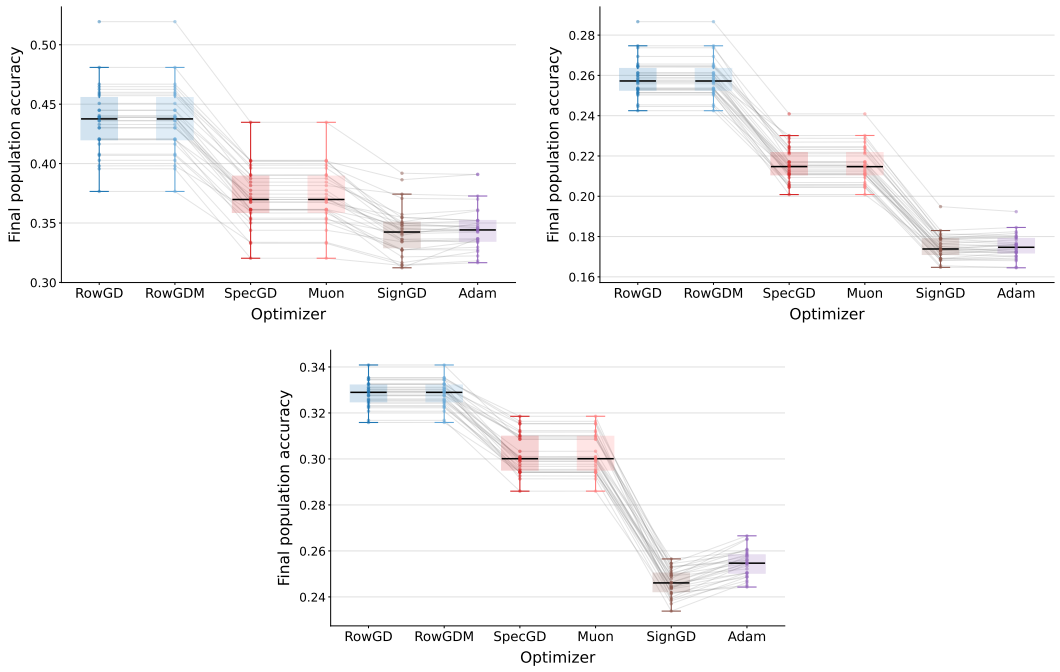

Figure 7: **Robustness to problem size and class-mean distribution.** Top: $(k, d) = (30, 20)$ (left) and $(120, 40)$ (right), compared with $(60, 20)$ in Figure 2 (left). Bottom: class means drawn independently and uniformly from the unit sphere instead of from $\mathcal{N}(\mathbf{0}, I_d/d)$.

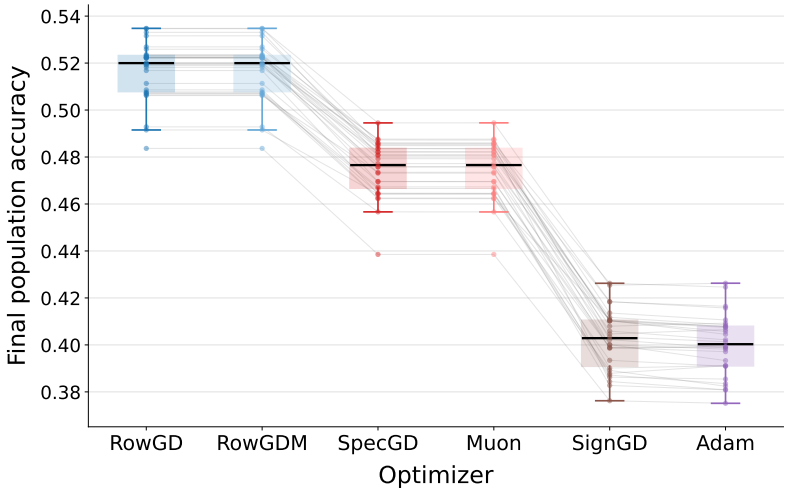

Figure 8: **Independent-orientation anisotropic experiment.** Representative setting with $s_\mu = 0.5$ and $s_\Sigma = 2$, where RowGD and RowGDM outperform SpecGD, Muon, SignGD, and Adam on all 30 paired seeds.

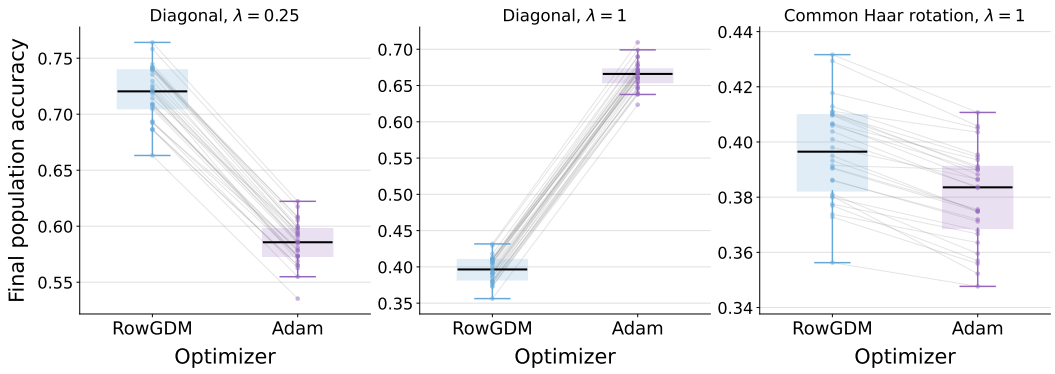

Figure 9: **Coordinate-aligned reversal and recovery.** RowGDM outperforms Adam at $\lambda = 0.25$, Adam outperforms RowGDM at $\lambda = 1$, and the RowGDM-over-Adam ordering is restored after applying the same Haar rotation.

H.2 Language Model Experiment

Model and starting checkpoint. We use a LLaMA-60M causal language model with 58.07M parameters exactly, consisting of 8 transformer layers with hidden size 512, intermediate size 1376, 8 attention heads, and vocabulary size $32{,}000$. The starting checkpoint is obtained by training the full model from scratch on C4 for $11{,}000$ updates with global batch size 512, sequence length 256, a peak learning rate of 0.005, cosine decay, and $1{,}100$ warmup steps. This pretraining uses a hybrid optimizer: Muon is applied to the two-dimensional attention and MLP projection matrices, while AdamW is used for the token embedding, language-model head, and one-dimensional parameters. The resulting checkpoint has validation perplexity 28.23501. For the last-layer experiment, we freeze the entire model body, discard the pretrained output-head weights, zero-initialize the 32000×512 output head, and train only this $16{,}384{,}000$-parameter matrix.

Last-layer training horizon. The main experiment trains the zero-initialized output head for $22{,}000$ updates in bfloat16 using distributed data parallel training on two NVIDIA RTX A6000 GPUs, with global batch size 512, sequence length 256, a cosine learning-rate schedule with $1{,}100$ warmup steps, and zero weight decay and gradient clipping. An $11{,}000$-update last-layer run processes 1.10B non-padding training tokens, or approximately 18.9 tokens per total model parameter, close to the conventional 20-tokens-per-parameter Chinchilla reference; the $22{,}000$-update horizon approximately doubles this token budget to 2.20B tokens. We use the longer horizon to examine a later-training regime in which transient optimization effects are reduced and the trajectories may be closer to the limiting-direction behavior studied theoretically, without assuming that directional convergence has occurred by $22{,}000$ updates. The Chinchilla comparison is only a token-budget reference, since the model body is frozen and only the output head is optimized in this experiment.

Horizon check. We also run a separate $11{,}000$-update comparison with its own cosine schedule ending at $11{,}000$ steps and learning rates selected independently at that horizon. The selected learning rates are 0.00625 for RowSGDM, 0.0045 for Muon, and 0.005 for Adam, yielding final monitoring-set perplexities 28.5871, 28.6373, and 28.5914, respectively. Thus, RowSGDM is already numerically best at approximately the Chinchilla token-budget reference, although its advantage over Adam is only 0.0152% in this single-seed comparison. The $22{,}000$-step experiment therefore probes whether the qualitative ordering persists and becomes clearer at a later training horizon rather than relying on the $11{,}000$-step difference alone.

Data and evaluation. We use the Hugging Face `allenai/c4` dataset in streaming mode and tokenize with `t5-base`, truncating and padding sequences to length 256. For each training seed, RowSGDM, Muon, and Adam receive the same shuffled stream and batch sequence. This streaming shuffle is not the epoch-wise permutation of a fixed finite dataset used by the random-reshuffling optimizers in Section 2. The validation trajectory in Figure 2 (right) is obtained from a monitoring set evaluated every $1{,}000$ updates and containing $7{,}625{,}745$ valid next-token targets. Final evaluation uses a separate set of $237{,}723$ C4 validation documents containing $46{,}186{,}882$ valid targets and having zero document-hash overlap with the monitoring set. The held-out set is not used for learning-rate selection, early stopping, or checkpoint selection. NLL is averaged over the exact number of valid non-padding shifted targets, and perplexity is its exponential.

Optimizers and learning rates. RowSGDM uses momentum 0.9, row-wise RMS normalization, multiplier 0.2, and numerical stabilizer $\epsilon = 10^{-8}$. Muon uses momentum 0.95, five Newton–Schulz iterations, multiplier 0.2, and zero weight decay. Adam uses $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and zero weight decay. These implementations are practical counterparts of the RowSGDM, exact-SVD Muon, and zero-stability-constant Adam considered in our theory. For the $22{,}000$-step experiment, learning rates are selected using the monitoring set: we sweep $\{0.004, 0.005, 0.006, 0.007\}$ for RowSGDM and Adam and $\{0.0025, 0.003, 0.0035, 0.004, 0.0045, 0.005, 0.0055, 0.006\}$ for Muon, selecting 0.005, 0.005, and 0.004, respectively.

Boundary-alignment diagnostic. We additionally test whether the performance ordering is accompanied by the boundary geometry predicted by our analysis. For each valid held-out next-token position, let $h_t \in \mathbb{R}^{512}$ denote the frozen hidden representation immediately before the output head and let y_t denote its next-token class. For class i, we estimate $\widehat{\mu}_i = n_i^{-1} \sum_{t:y_t=i} h_t$ and retain the

Optimizer	NLL ↓	PPL ↓	$\widehat{D}_{\text{avg-bdry}}$ ↓	$\widehat{C}_{\text{bdry}}$ ↑
RowSGDM	$\mathbf{3.35503 \pm 0.00005}$	$\mathbf{28.6466 \pm 0.0013}$	$\mathbf{1.13194 \pm 0.00004}$	$\mathbf{0.35652 \pm 0.00005}$
Muon	3.35699 ± 0.00004	28.7027 ± 0.0012	1.13807 ± 0.00006	0.35018 ± 0.00007
Adam	3.35759 ± 0.00002	28.7199 ± 0.0006	1.13350 ± 0.00008	0.35481 ± 0.00009

Table 1: Held-out performance and empirical boundary alignment for last-layer-only LLaMA training. Values are mean ± sample standard deviation across five paired training seeds.

5,093 classes with $n_i \geq 1000$. We uniformly sample a fixed set $\mathcal{P}$ of 1,000,000 unordered class pairs and use the same pairs for every optimizer and seed. For a trained output head with rows $w_i^\top$, define

$$\widehat{u}_{ij}^W := \frac{w_i - w_j}{\|w_i - w_j\|_2}, \qquad \widehat{u}_{ij}^\mu := \frac{\widehat{\mu}_i - \widehat{\mu}_j}{\|\widehat{\mu}_i - \widehat{\mu}_j\|_2},$$

and measure

$$\widehat{D}_{\text{avg-bdry}}(W) := \frac{1}{|\mathcal{P}|} \sum_{\{i,j\} \in \mathcal{P}} \|\widehat{u}_{ij}^W - \widehat{u}_{ij}^\mu\|_2, \qquad \widehat{C}_{\text{bdry}}(W) := \frac{1}{|\mathcal{P}|} \sum_{\{i,j\} \in \mathcal{P}} \langle \widehat{u}_{ij}^W, \widehat{u}_{ij}^\mu \rangle.$$

Lower $\widehat{D}_{\text{avg-bdry}}$ and higher $\widehat{C}_{\text{bdry}}$ indicate closer alignment between learned classifier boundaries and empirical class-mean differences.

Results. Consistent with the final validation-perplexity ordering in Figure 2 (right), RowSGDM achieves lower held-out NLL and perplexity than both Muon and Adam on each of the five paired seeds. RowSGDM also has the smallest boundary discrepancy and the largest boundary cosine on every seed, so its performance advantage is accompanied by closer alignment between learned classifier boundaries and empirical class-mean differences. Table 1 summarizes the five-seed held-out results.

The mean paired RowSGDM-minus-Muon NLL difference is -0.001957, with a descriptive 95% Student-t interval $[-0.002058, -0.001857]$, and the RowSGDM-minus-Adam difference is -0.002555, with interval $[-0.002621, -0.002489]$. The mean reductions in $\widehat{D}_{\text{avg-bdry}}$ are 0.006137 relative to Muon and 0.001558 relative to Adam. Together, the validation, held-out, and boundary-alignment results are consistent with the qualitative ordering and geometric mechanism predicted by our theory.

LLaMA-130M experiment. We additionally repeat the last-layer comparison on a separately C4-pretrained LLaMA-130M backbone, whose actual parameter count is 134.1M, with 12 transformer layers and hidden size 768. The tokenizer, sequence length, global batch size, evaluation sets, optimizer definitions, and five paired seeds are the same as above. We train the zero-initialized $32{,}000 \times 768$ output head for $55{,}000$ updates, with each selected run processing approximately 5.47B valid next-token targets, and a common full-horizon learning-rate sweep selects 0.002828 for all three optimizers. RowSGDM attains the lowest held-out perplexity on every seed, with five-seed mean ± sample standard deviation 22.3797 ± 0.0014, compared with 22.4096 ± 0.0030 for Muon and 22.4153 ± 0.0010 for Adam. RowSGDM also yields the smallest boundary discrepancy on every seed, with mean $\widehat{D}_{\text{avg-bdry}} = 1.19051$, compared with 1.19664 for Adam and 1.20028 for Muon, providing a larger-model check of the same qualitative advantage.

H.3 Spectral Exponent for Language Model Representation

We empirically examine whether a fitted class-mean spectral exponent below one can occur in language-model representations, as in the regime $s_\mu < 1$ of Theorem 4.2. We also measure class-conditional residual spectral exponents as descriptive statistics of within-class anisotropy. These measurements are finite-dimensional spectral summaries and are not intended as a direct test of the full probabilistic model underlying the theorem.

Experimental setup. We analyze final pre-head representations from OLMoE-1B-7B-0125 (Muennighoff et al., 2025), Qwen3-30B-A3B (Yang et al., 2025), and AMUSE-720M (Kim et al., 2026b),

Table 2: Full-rank pointwise-OLS spectral exponents at the three final checkpoints. K is the number of observed classes used for the class-mean covariance, and K_Σ is the number of reliable classes used for the residual summary.

Model	K	$\widehat{s}_\mu$	K_Σ	Mean $\widehat{s}_{\Sigma,i}$	Median	Max
OLMoE-1B-7B-0125	47,190	0.957735	218	0.987339	0.973537	1.374366
Qwen3-30B-A3B	72,577	0.932780	120	1.057924	1.045483	1.258163
AMUSE-720M	49,348	0.985366	211	1.054793	1.045965	1.402045

all with hidden dimension $d = 2048$. For each model, we collect exactly $20{,}000{,}000$ valid same-document next-token positions from the English validation split of C4. At a valid position, let $h_{it} \in \mathbb{R}^d$ denote the final normalized hidden representation when the next-token class is i, and let n_i be the number of observed positions from class i. We define the empirical class mean by

$$\widehat{\mu}_i := \frac{1}{n_i} \sum_{t=1}^{n_i} h_{it}.$$

Class-mean spectral exponent. Our primary mean-spectrum analysis assigns equal weight to every observed next-token class. Let

$$\mathcal{I}_\mu := \{i : n_i \geq 1\}, \qquad K := |\mathcal{I}_\mu|,$$

and define

$$\widehat{\bar{\mu}} := \frac{1}{K} \sum_{i \in \mathcal{I}_\mu} \widehat{\mu}_i, \qquad \widehat{\Gamma} := \frac{1}{K} \sum_{i \in \mathcal{I}_\mu} (\widehat{\mu}_i - \widehat{\bar{\mu}})(\widehat{\mu}_i - \widehat{\bar{\mu}})^\top.$$

Thus $\widehat{\Gamma}$ is class balanced rather than token-frequency weighted; in particular, every observed class contributes one empirical class mean. If $\widehat{\lambda}^\mu_1 \geq \cdots \geq \widehat{\lambda}^\mu_{2048} > 0$ are the eigenvalues of $\widehat{\Gamma}$, we estimate the mean exponent by the unweighted pointwise OLS fit

$$\log \widehat{\lambda}^\mu_r = a - \widehat{s}_\mu \log r + \varepsilon_r, \qquad r = 1, \ldots, 2048.$$

Each integer rank therefore contributes one regression observation; the fit is not log binned or rank weighted.

Residual spectral exponents. For classes with sufficiently many observations, we form the within-class sample covariance

$$\widehat{\Sigma}_i := \frac{1}{n_i - 1} \sum_{t=1}^{n_i} (h_{it} - \widehat{\mu}_i)(h_{it} - \widehat{\mu}_i)^\top$$

and fit

$$\log \widehat{\lambda}^\Sigma_{i,r} = a_i - \widehat{s}_{\Sigma,i} \log r + e_{i,r}, \qquad r = 1, \ldots, 2048.$$

The reliable-class threshold is $n_i \geq 8192$ for OLMoE and AMUSE and $n_i \geq 16384$ for Qwen3. This leaves 218, 211, and 120 residual classes, respectively.

Results. For the three final checkpoints examined here, the primary class-mean point estimates in Table 2 all satisfy $\widehat{s}_\mu < 1$: 0.957735 for OLMoE, 0.932780 for Qwen3, and 0.985366 for AMUSE. The residual exponents are more heterogeneous and can exceed one, as reflected by the maxima in the table.

OLMoE across annealing. We additionally apply the $\widehat{s}_\mu$ estimator to six published OLMoE snapshots spanning 0–101B added annealing tokens. The resulting estimates are 0.956727, 0.957792, 0.956569, 0.956530, 0.958359, and 0.957735. Thus, every sampled estimate is below one, the total range is only 0.001830, and there is no monotone trend across the sampled snapshots.